Prototipo de plataforma inteligente de bajo costo orientada a la nube basada en inteligencia de negocios para la mejora de toma de decisiones de PYMES

Cabrera Teanga, Diego Fernando y Sosa Cañar, Alexander Francisco

Departamento de Ciencias de la Computación

Carrera de Ingeniería de Sistemas e Informática

Trabajo de titulación, previo a la obtención del título de Ingeniero en Sistemas e Informática

Ing. Marcillo Parra, Diego Miguel

19 de febrero del 2020

URKUND

Urkund Analysis Result

Analysed Document:	Tesis Cabrera-Sosa.docx (D63480573)
Submitted:	2/5/2020 4:05:00 PM
Submitted By:	${Xml.Encode(Model.Document.Submitter.Email)}
Significance:	0 %

Sources included in the report:

Instances where selected sources appear:

0

Ing. Marcillo Parra, Diego Miguel

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**Prototipo de plataforma inteligente de bajo costo orientada a la nube basada en inteligencia de negocios para la mejora de toma de decisiones de PYMES**" fue realizado por los señores **Cabrera Teanga, Diego Fernando y Sosa Cañar, Alexander Francisco,** el cual ha sido revisado y analizado en su totalidad por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí. 13 de febrero de 2020

Firma:

Firmado electrónicamente por:
DIEGO MIGUEL
MARCILLO
PARRA

Ing. Marcillo Parra, Diego Miguel

C. C 1710802925

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

RESPONSABILIDAD DE AUTORÍA

Nosotros, **Cabrera Teanga, Diego Fernando y Sosa Cañar, Alexander Francisco**, con cédulas de ciudadanía N° 1723352207 y 1715037410, declaramos que el contenido, ideas y criterios del trabajo de titulación**: Prototipo de plataforma inteligente de bajo costo orientada a la nube basada en inteligencia de negocios para la mejora de toma de decisiones de PYMES** es de nuestra autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos, y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas

Sangolquí, 13 de febrero de 2020

Firmas:

Cabrera Teanga, Diego Fernando

C.C.: 1723352207

Sosa Cañar, Alexander Francisco

C.C.: 1715037410

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

AUTORIZACIÓN DE PUBLICACIÓN

Nosotros **Cabrera Teanga Diego Fernando y Sosa Cañar Alexander Francisco,** con cédulas de ciudadanía N° 1723352207 y 1715037410, autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **Prototipo de plataforma inteligente de bajo costo orientada a la nube basada en inteligencia de negocios para la mejora de toma de decisiones de PYMES** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de nuestra responsabilidad.

Sangolquí, 13 de febrero de 2020

Firmas:

Cabrera Teanga, Diego Fernando

C.C.: 1723352207

Sosa Cañar, Alexander Francisco

C.C.: 1715037410

Dedicatoria

Dedicó este proyecto a mi familia, amigos, son los seres más queridos en mi vida que sin el apoyo y paciencia de todos ellos esto no fuera posible. A aquellos que no dudaron de mí y siempre me apoyaron con sus palabras de cordura y a mis maestros de la vida universitaria.

Este proyecto va dedicado a familiares, maestros, amigos y todos aquellos que de forma directa o indirecta fueron parte de esta experiencia académica.

Agradecimientos

Agradezco a mi madre y a mi padre, siempre me apoyaron en toda mi vida para que pueda tener una buena educación, lo que tengo y he logrado hasta hoy en día, es gracias a ellos, lo más valioso que me pudo pasar es tenerlos como padres.

Agradezco especialmente a mis abuelos, a mi padre y madre, cuyo apoyo y confianza fueron la principal motivación para llevar a cabo este proyecto. Sin su incondicional apoyo nada de esto hubiese sido posible.

Tabla de Contenidos

Índice de tablas

Índice de figuras

Resumen

Cada año el negocio de pequeñas y medianas empresas crece junto con la información que estas generan. El objetivo de volverse más competentes, abarcar mayor terreno en el mercado y anticiparse a sus competidores hace que las PYMEs busquen darle una utilidad a la información que poseen a través de la implementación de una solución de Inteligencia de Negocios, que las ayude a cumplir estos objetivos por medio de toma de decisiones. Implementar este tipo de soluciones conlleva una serie de gastos que las organizaciones deben afrontar, tales como: infraestructura adecuada para el procesamiento de datos, licencias de software, contratación de expertos en análisis de datos, capacitación al personal, etc.

Este trabajo propuso diseñar e implantar un prototipo de plataforma que brinde soluciones de Inteligencia de Negocios, por medio de herramientas de código abierto basadas en la nube para ayudar a la toma de decisiones, reduciendo los costos de implementación.

Al poner a prueba este prototipo, se concluyó en que los resultados obtenidos fueron favorables, tanto en la capacidad que demostró tener para procesar enormes cantidades de datos de diferentes fuentes y en diferentes formatos, así como en la facilidad de uso que presenta al usuario para dar uso e interpretación a los mismos; reduciendo significativamente el costo de su implementación.

PALABRAS CLAVE:

- **INTELIGENCIA DE NEGOCIOS**
- **PYMES**
- **ANALISIS DE DATOS**

Abstract

Every year the business of small and medium enterprises grows along with the information they generate. The objective of becoming more competent, covering more market and anticipating its competitors means that PYMEs seek to give a usefulness to the information they have through the implementation of a Business Intelligence solution, which helps them fulfill these objectives through decision making. Implementing these kind of solutions entails a series of expenses that organizations must face, such as: adequate infrastructure for data processing, software licenses, hiring experts in data analysis, staff training, etc.

This work proposed to design and implement a prototype platform that provides Business Intelligence solutions, through cloud-based open-source tools to help decision making, reducing implementation costs.

When testing this prototype, it was concluded that the results obtained were favorable, both in the capacity it proved to have to process huge amounts of data from different sources and in different formats, as well as in the ease of use it presents to the user for give use and interpretation to them; significantly reducing the cost of its implementation.

KEY WORDS:

- **BUSINESS INTELLIGENCE**
- **PYMES**
- **DATA ANALYSIS**

Capítulo 1 - Introducción

Antecedentes

Desde hace mucho tiempo los datos no aportaban ningún valor relevante, los datos deben ser analizados para que se transformen en información que sea relevante para la toma de decisiones o aportar al conocimiento de algo específico.

Las grandes cantidades de datos debían pasar por un análisis que tardaban mucho tiempo para poder obtener resultados que ayuden a la toma de decisiones, es por ello que la necesidad hizo que nuevas maneras de realizar análisis aparecieran, como se comenta a continuación.

El termino Business Intelligence (BI)[1] se lo usa para describir como una organización se puede beneficiar de la información al poder reunirla y actuar sobre ella antes que su competencia (Negash & Gray, 2008).

La tecnología no era considerada como una herramienta que ayude a mejorar los temas de asuntos comerciales entre empresas hasta inicios del siglo XX. El artículo escrito por el científico (Luhn, 1958), describía un sistema automático que tenía por objetivo difundir información a diferentes departamentos de cualquier tipo de organización de tipo científica, industrial o gubernamental, tal fue el impacto de este artículo que varios trabajos se realizaron posteriormente referente a este tema.

Por otro lado, desde el momento en que surgieron las primeras computadoras, medios de almacenamiento, dispositivos móviles y centros de datos modernos,

[1] Inteligencia de Negocios o en inglés Business Inteligence (BI): es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios.

hemos generado y recopilado datos e información a gran escala. Con este aumento masivo de datos, resulta complejo realizar búsquedas o análisis y transformar los datos en información que sea útil para la toma de decisiones en las empresas y por ello han surgido en el mercado un sin número de herramientas que brindan facilidades para todo el tratamiento de la información, sin embargo, el costo de estas es elevado lo que lo vuelve accesible únicamente para grandes corporaciones u empresas con un gran capital.

Por otra parte, las empresas durante años han usado la informática y los sistemas que se vienen desarrollando como herramientas operativas con el fin de simplificar procesos. Hoy en día, esta forma de ver a las Tecnologías de la Información debe cambiar, porque no debemos pretender usarlas tan solo como instrumentos para optimizar costos.

Planteamiento del Problema

Las PYMEs[2] buscan utilizar la información que poseen para implementar una solución que les ayude a mejorar sus procesos de producción, volverlos más competentes, y optimizar la toma de decisiones.

Al no tener una noción clara del valor que posee esta información y de cómo tratarla para que ayude a alcanzar los objetivos de la organización, las PYMES se ven orilladas a contratar o depender de terceros que los ayuden a implementar una solución de Inteligencia de Negocios.

[2] Empresa pequeña o mediana en cuanto a volumen de ingresos, valor del patrimonio y número de trabajadores.

Es aquí en dónde surge el problema en cuestión, ya que adquirir los servicios de consultores o expertos en el manejo e interpretación de datos resulta sumamente costoso. En muchas ocasiones las empresas se ven obligadas a comprar una nueva infraestructura en la que se pueda alojar la plataforma de Inteligencia de Negocios, la misma que deberá contar con potentes equipos capaces de obtener, procesar e interpretar enormes flujos de datos históricos y en tiempo real.

Otro gran inconveniente al que las organizaciones se deben enfrentar, una vez tengan una solución de Inteligencia de Negocios, tiene que ver con el talento humano de la propia empresa, éste debe tener los conocimientos suficientes y necesarios en administración de la plataforma de Inteligencia de Negocios, así como de experticia en el manejo de las diferentes fuentes de datos que maneja la empresa y sobre todo tener claro el giro de negocio de la organización y los factores claves para cumplir los objetivos de la misma. Esto conlleva a otro gasto significativo para la compañía ya que debe elegir entre contratar profesionales en Inteligencia de Negocios y Análisis de Datos o capacitar en estos temas al personal que corresponda; ambas opciones representan más gastos.

Justificación

Debido a que en la actualidad no hay muchos científicos de datos que puedan realizar análisis de los datos y obtener resultados óptimos, con este proyecto vamos a ayudar a que este grupo reducido crezca al nosotros inteligenciarnos en esta área.

La toma de decisiones basadas en datos históricos y actualizados en tiempo real, son de gran importancia a nivel gerencial por lo que Inteligencia de Negocios facilita este proceso gracias a las bondades que proporciona:

- No requiere de personal altamente capacitado.
- El prototipo de plataforma se puede armar dinámicamente, tan solo se debe tener el conocimiento del proceso para poder estructurar las consultas.
- Se puede tomar decisiones en menor tiempo gracias a que el análisis se lo realiza en tiempo real.
- Los sectores de negocio en los que se puede adaptar el prototipo de plataforma son diversos gracias a que la Inteligencia de Negocios es independiente del giro del negocio[3] de cada empresa.
- Al poder alojar en la nueve se puede reducir el gasto que se debería afrontar para adquirir una infraestructura propia para la plataforma. Transformando a la aplicación en un servicio al que las organizaciones podrán acceder por medio de una conexión a internet (SaaS)[4].

Para mejorar la toma de decisiones a nivel gerencial y en menor tiempo se propone desarrollar e implantar una plataforma inteligente bajo costo para la toma de decisiones basadas en Inteligencia de Negocios.

[3] El giro de una empresa se refiere a la actividad o negocio que desarrolla la misma.
[4] Saas: Un proveedor de servicios ofrece acceso a un entorno basado en la nueve en el cual los usuarios pueden acceder a aplicaciones, el proveedor proporciona la infraestructura subyacente.

Objetivos

Objetivo General

Desarrollar un prototipo de plataforma inteligente de bajo costo a través de internet, mediante herramientas y metodologías de Inteligencia de Negocios para análisis de datos, que permita mejorar y optimizar la toma de decisiones estratégicas de una PYME en base a datos históricos y generados en tiempo real (streaming).

Objetivos Específicos

Investigar herramientas y metodologías para el desarrollo de soluciones de Inteligencia de Negocios mediante una revisión sistémica de literatura que permitan plantear el prototipo de la plataforma indicada.

Diseñar la arquitectura del prototipo orientado a internet, utilizando la metodología seleccionada de la revisión sistémica de literatura.

Implantar el prototipo, mediante el uso de las herramientas de código abierto resultado de la investigación.

Validar el funcionamiento del prototipo con una PYME, obteniendo datos históricos y generados en tiempo real, para seleccionarlos, transformarlos y procesarlos aplicando conceptos de Ciencia de Datos (Data Science), para finalmente presentar resultados de forma intuitiva y amigable al usuario.

Alcance

Desarrollar e implantar un prototipo de plataforma inteligente de bajo costo basada en Inteligencia de Negocios para poder mejorar la toma de decisiones en diversos giros del negocio que contará con las siguientes características:

- Será desarrollada con herramientas de código abierto, lo que lleva a prescindir de gastos en licencias de aplicaciones o programas propietarios sin dejar de lado la calidad del producto.
- Va a ser alojada en la nube.
- Plataforma amigable, permitiendo que el usuario final sea capaz de crear sus propios cuadros de mando, gráficos estadísticos, PKI's[5], etc., sin tener que ser un experto en la materia; bastará con que posea el conocimiento suficiente y necesario para que logre obtener el resultado esperado.
- Procesamiento en tiempo real, logrando que los datos generados en streaming sean procesados inmediatamente y puedan ser interpretados y mostrados al instante para que la toma de decisiones sea rápida y acertada.
- Los datos serán almacenados de una base de datos no relacional (NoSql)[6] para el mejor manejo de datos no estructurados.
- Implantar la plataforma para evidenciar la eficacia y veracidad de los resultados del análisis.

Hipótesis

H_0: La plataforma inteligente de bajo costo orientada a la nube permitirá implementar soluciones de Inteligencia de Negocios, reduciendo significativamente los costos en comparación con aplicaciones propietarias.

[5] Indicador clave o medidor de desempeño o indicador clave de rendimiento, es una medida del nivel del rendimiento de un proceso.
[6] NoSQL: es una amplia clase de sistemas de gestión de bases de datos que difieren del modelo clásico entidad relación.

H_1: La plataforma inteligente de bajo costo orientada a la nube permitirá implementar soluciones de Inteligencia de Negocios, sin reducir significativamente los costos en comparación con aplicaciones propietarias.

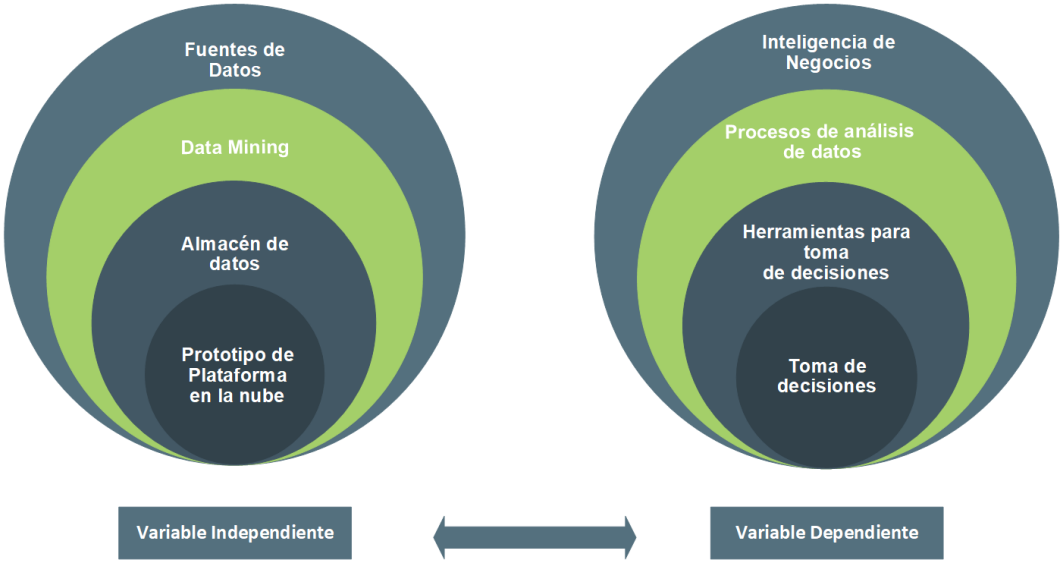
Categorización de las variables de investigación

Variable Independiente: Prototipo de plataforma en la nube para Inteligencia de Negocios.

Variable Dependiente: Reducción de costos y mejora en el manejo de información para la toma de decisiones.

Figura 1

Variables de investigación



Definición de la Investigación

Para llevar a cabo la investigación se ha establecido una metodología específica para el desarrollo de este proyecto que está basada en diferentes fases que buscan obtener un entendimiento del problema y necesidades de una PYME y con el producto final lograr cumplir los objetivos de esta. Se han definido las siguientes fases:

a. FASE I – Investigación: En esta fase lo que se realiza es un estudio de la problemática en su totalidad. En el caso de este proyecto, se trata del estudio de las reglas de negocio de una PYME y los datos que esta posee. Se debe determinar qué información va a ser de utilidad para lograr los objetivos de una PYME. A su vez, también se debe investigar sobre como las herramientas orientadas a la nube pueden ser de utilidad para la conexión con estas fuentes de datos.
b. FASE II – Diseño: Con los resultados de la fase anterior, se diseña un prototipo que pueda ajustarse a las reglas de negocio de una PYME en cuestión. Para esto es necesario extraer, transformar y cargar los datos de las fuentes seleccionadas para quedarnos únicamente con la información que genere ayude a la toma de decisiones y en un único formato.
c. FASE III – Implementación: Con el prototipo ya diseñado, es necesario implementarlo en un ambiente de producción (en la nube) y con información real. Se consideran todas las especificaciones y necesidades de una PYME y se establece un entorno en dónde el modelo puede ser puesto a prueba de forma controlada.
d. FASE IV – Validación: Se realizan pruebas del prototipo montado en la nube para validar su funcionamiento, usabilidad y disponibilidad. En esta fase se obtienen los resultados de todas las fases anteriores, validando al final si el prototipo pudo ser aplicado adecuadamente al giro de negocio de una PYME, mejorando y optimizando la toma de decisiones.

Capítulo 2 – Marco teórico y estado del arte

En la actualidad estamos rodeados constantemente de grandes cantidades de datos que crecen excesivamente. Estos datos al ser analizados y estudiados pueden otorgar información realmente valiosa para una persona o empresa. El problema para analizar adecuadamente estos datos es que requerimos tecnologías poco convencionales que han sido desarrolladas específicamente para dar solución al análisis de grandes cantidades de datos.

Big Data

Es un término que se usa frecuentemente en la actualidad para referirse a grandes volúmenes de datos. Dichos datos pueden ser estructurados y no estructurados. También pueden presentarse en varios tipos de formatos como por ejemplo: .txt, .mp3, .mp4, .json, etc. Estos datos pueden estar almacenados en discos duros o ser tomados en tiempo real vía streaming.

El Big Data está revolucionando la forma en que las empresas funcionan al permitir que grandes cantidades de datos puedan ser analizados para obtener resultados que ayuden a tomar mejores decisiones y movimientos estratégicos a la empresa.

Cada documento sobre las técnicas de procesamiento de información innovadora y rentable ha sido desarrollado en código abierto dentro del ecosistema cada vez más grande llamado Big Data.

Hoy en día, tenemos varias herramientas que facilitan la recolección, almacenamiento y procesamiento de información a gran escala. Las características importantes que posee Big Data son conocidas como las 5Vs:

Característica

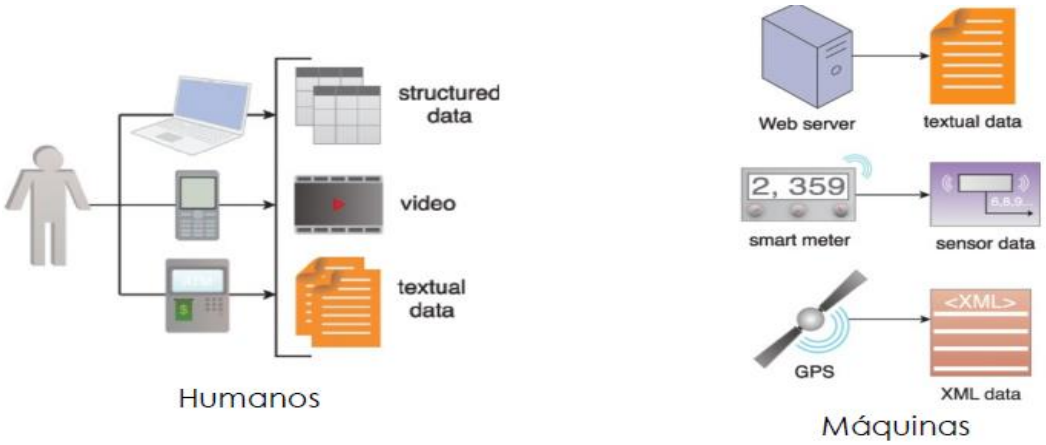
Volumen: Según (IBM, 2016), se generan 2.5 quintillones de bytes de datos continuamente, con esto se quiere decir que en los últimos dos años sea generando el 90% de los datos que existen en el mundo hasta la fecha. Por este motivo el volumen de datos que existe para analizar es realmente enorme tornándose un tanto complejo.

Velocidad: En cualquier aspecto de tecnología la velocidad juega un rol muy importante. Para el análisis del tiempo real es esencial la velocidad en cualquier empresa, ocasionalmente las empresas necesitan tomar decisiones en tiempos relativamente cortos lo que conlleva a que el flujo de la información sea rápido.

Variedad: Debido a los diferentes tipos de datos que se generan contantemente por los diversos dispositivos electrónicos, Big Data es capaz de procesar y analizar esta información sin dejar de lado ningún tipo de formato sea audio, video, texto etc. Tal como se puede apreciar en la Figura 1. En resumen, todos estos datos son estructurados (creados en base a un modelo de datos o esquema por ejemplo datos de una base de datos relacional), no estructurados (no están creados en base a un modelo de datos o esquema por ejemplo binarios, audio y video) y semiestructurados (presentan un nivel de estructura y consistencia, pero no de manera relacional por ejemplo archivos xml, json y datos de sensores). En muchas ocasiones hay que mezclar diversos tipos de formatos para poder analizarlos y obtener información importante.

Figura 2

Fuentes de tipos de datos



Con el paso del tiempo estas características no fueron suficientes para describir todo lo que Big Data es capaz de hacer para el almacenamiento, procesamiento y análisis de información apareciendo dos características más las cuales son:

Veracidad: Debido al volumen y variedad de datos que Big Data es capaz de procesar, existe mayor incertidumbre en los datos ya que se debe determinar qué datos son de utilidad y cuales son datos basura o que se pueden marginar.

Valor: Todo el procesamiento y análisis que se realiza sobre los datos, debe generar valor para la empresa caso contrario Big Data no estaría aportando en nada a la misma.

En esta tesis las características mencionadas son importantes para poder realizar el análisis sobre los datos obtenidos mediante streaming y poder generar valor sobre los mismos. Determinar cuál es la mejor técnica de Big Data para realizar este análisis dará como resultado un mejor producto final.

Empresas de Renombre

- **Facebook**

A causa de los más de 960 millones de usuarios que posee esta red social, el crecimiento masivo de la información se acelera, toda esta información Facebook la almacena en bases de datos relacionales y no relacionales robustas para soportar millones de consultas que se realizan por día alrededor del mundo. Todo lo que hacemos en esta red social queda almacenado como lo son los clics que hacemos en la publicidad, notificación, me gusta, comentarios, mensajes enlaces, fotos, etc. Generan datos que son de valor para que la empresa pueda realizar un seguimiento de varios registros.

- **Twitter**

Es considerada la segunda red social más grande, pero genera menos datos en comparación con la aplicación de citas como Tinder. Por otro lado, los usuarios de twitter generan millones de Tweets por hora que son almacenados para poder ser analizados.

- **YouTube**

El audio y video también pueden ser analizados para poder conocer tendencias o preferencias en Internet. YouTube es el gigante del video alrededor del mundo, se dice que cada minuto los usuarios están cargando más de 300 horas video.

Cifras y Hechos

El 91% de las personas que lideran las grandes marcas usan datos de los clientes para poder tomar decisiones del negocio.

El porcentaje total de los datos totales del mundo se ha creado justo en los últimos dos años es del 90%.

En el mundo el 87% de las empresas están de acuerdo en capturar y compartir los datos correctos, es importante para medir eficazmente el Retorno de la Inversión (ROI) en su propia empresa.

En Facebook 30 mil millones contenido se comparten por los usuarios en cada mes.

Ciencia de Datos

La Ciencia de Datos en inglés Data Science, es el estudio que se encarga de analizar los datos, su objetivo es obtener conocimiento valedero de manera automatizada desde un conjunto de datos provisto. Mediante la ciencia de datos está aportando al desarrollo de nuevas e innovadoras ramas de la ciencia en el ámbito de las humanidades o ciencias sociales.

Al combinar varias ciencias como la estadística, matemáticas, informática da como resultado el poder obtener grandes cantidades de información que no podríamos obtener antes sin un estudio exhaustivo de los datos. La metodología que emplea un científico de datos requiere comprender varios puntos, en gran parte del trabajo para realizar un análisis de datos eficiente consiste en un tratamiento previo de los datos (preprocesamiento).

Con frecuencia las fuentes de información están desordenadas e incompletas mucho más en grandes cantidades de datos. Para preprocesar los datos se debe utilizar hardware y software especializado. Una vez que los datos han sido preprocesados podemos trabajar sobre ellos de manera eficiente. Luego se debe iterar varias veces sobre ellos para analizarlos y poder conseguir un modelo de predicción apropiado.

Para poder conseguir un modelo de predicción apropiado hay que analizar varias veces no se puede conseguirlo en la primera ocasión, este proceso requiere de mucho conocimiento y experimentación. Él científico de datos además de construir modelos tiene también el objetivo de facilitar estos modelos de predicción a personas ajenas a la ciencia de datos para obtener valor en dichos datos. Es decir que no sirve de nada desarrollar un programa que provea patrones a partir de datos, si no se obtiene valor o una aplicación en producción que aporte a la empresa.

Características

- Conocer del tema en cuestión a tratar.
- Conocer de las tecnologías de la información.
- Dominar matemáticas y estadísticas a buen nivel.

Business Intelligence

Definición

Business Intelligence es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar la forma en que actualmente se toma decisiones en las PYMEs.

Business Intelligence desempeña un factor relevante y estratégico en las PYMEs o grandes organizaciones, porque genera una ventaja potencial y a la vez competitiva, ya que proporciona información clave para afrontar los problemas de negocio (Curto Díaz, 2010).

Características

- **Accesibilidad a la información**: Garantizar el acceso de los usuarios a todos los datos sin que influyan la procedencia de estos.
- **Apoyo en la toma de decisiones**: Se debe permitir al usuario que seleccione y manipule los datos que le interesen para realizar un análisis personalizado que ayude a identificar las fortalezas y debilidades de su actividad.
- **Orientación al usuario final**: Será clave que el sistema sea intuitivo y de fácil manejo para que no importen demasiado los conocimientos técnicos.

Bases de datos

Dependiendo el giro del negocio que desempeñe cada empresa, tendrá que realizar un análisis exhaustivo para determinar porque tipo de base de datos optar, de esta decisión dependerá que la empresa pueda dar solución a un requerimiento que el cliente solicite.

Relacionales

Las bases de datos relacionales son las que se manejan en los sistemas tradicionales, que se podía realizar un CRUD para poder dar mantenimiento a la información de una determinada empresa. Está formada de elementos de datos

que poseen relaciones predefinidas, conocidos como tablas que están conformadas por columnas y filas.

- Tabla: Específicamente son utilizadas para guardar información en objetos que a su vez van a ser representados en la base de datos.
- Columna: Una tabla puede estar formada por varias columnas, cada una de ellas posee un determinado tipo de datos.
- Fila: Se podría decir que una fila posee un identificador único conocido como clave primaria, las filas de otras tablas pueden relacionarse por su clave foránea.

Características importantes

- **SQL**: Es el lenguaje de conosulta que se utiliza en las interfaces principales de las bases de datos relacionales sus siglas significan Structure Query Languaje.
- **Integridad de los datos**: Utilizan restricciones para la integridad de los datos en la base de datos.
- **Transacciones**: Se considera que una o más sentencias SQL son una transacción de base de datos, las sentencias son una secuencia de operaciones que llegan a conformar una unidad lógica de trabajo.
- **ACID:** Para garantizar que las transacciones tengan integridad en sus datos, cada transacción debe cumplir con ser; atómicas, coherentes, aisladas y duraderas (ACID).

Las bases de datos relacionales más conocidas y usadas hoy en día son:

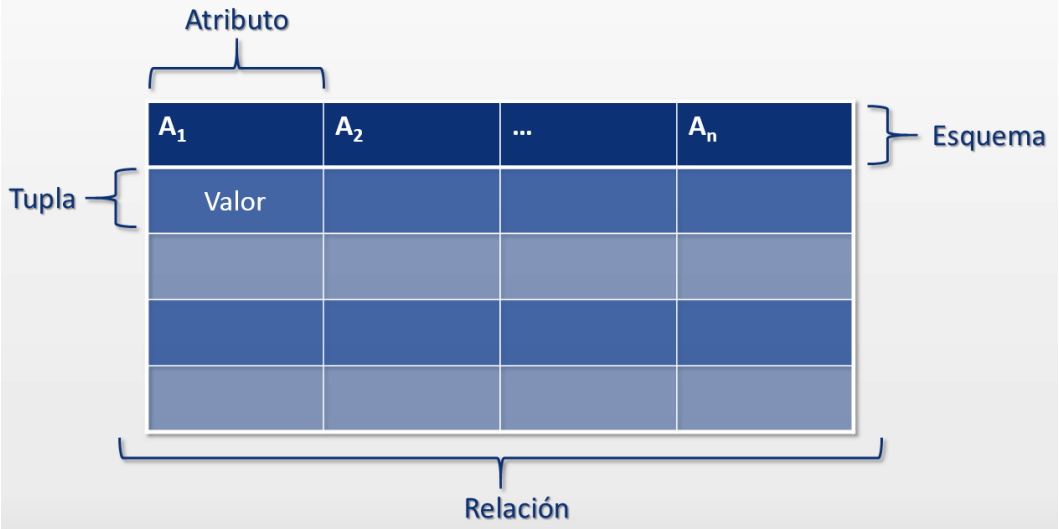
- Oracle

- SQL Server
- PostgreSql
- MySql
- MariaDB
- Sybase

A los datos de puede acceder de varias maneras sin tener que hacer una reestructuración de las tablas de la base de datos (AWS, 2019).

Figura 3

Elementos de una Tabla de base de datos relacional



Nota. (Guide, 2020)

No Relacionales NoSQL

Estas bases de datos se diseñan para modelos de datos específicos y tienen esquemas flexibles que se usan en aplicaciones modernas. Estas bases en los últimos años han llegado a darse a conocer porque son fáciles de desarrollar, su funcionalidad y el rendimiento a escala.

Son capaces de soportar una gran variedad de modelos de datos, que pueden ser datos estructurados o no estructurados como pueden ser documentos, gráficos, video, audio, etc. (AWS, Amazon Web Services, 2019).

Figura 4

Base de datos NoSQL

Como funcionan

Estas bases al usar gran variedad de modelos de datos están optimizadas para aplicaciones que requieren procesar grandes volúmenes de datos reduciendo la latencia y con el uso de modelos de datos flexibles, generalmente se indexa y se almacena en un documento suelen manear archivos de tipo JSON.

Herramientas

Propietarias

A continuación, se presenta un listado de algunas de las herramientas propietarias más utilizadas para la implementación de una solución de Inteligencia de Negocios, destacando los montos que los usuarios deben invertir para poder utilizarlas.

Tableau

Tabla 1

Tableau descripción

Versión	Información del producto	Vigencia del contrato	Precio por usuario
Tableau Desktop Personal Edition	Solución de Visualización y Análisis para datos almacenados en archivos (100 GB)	1 año	$999,00
Tableau Desktop Professional Edition	Solución de Visualización y Análisis para cualquier dato	1 año	$1999,00
Tableau Online (Hosted version of Tableau Server)	Versión Hospedada de Tableau Server	1 año	$500,00

MicroStrategy

Tabla 2

MicroStrategy descripción

Versión	Información del producto	Vigencia del contrato	Precio por usuario
DESKTOP	Solución de Visualización y Análisis para datos almacenados en archivos	1 año	$600,00
WEB	Solución de Visualización y Análisis para cualquier dato	1 año	$600,00
MOBILE	Disponible en las tiendas Itunes y Google Play	1 año	$600,00

ARCHITECT	Un conjunto de herramientas de desarrollo y migración que automatizan procesos, ahorran tiempo y administran la aplicación durante su ciclo de vida.	1 año	$5000,00
SERVER	Permite la conexión a fuentes de datos múltiples incluye herramientas de administración y supervisión	1 año	$1200,00
CLOUD MicroStrategy	Amazon Web Services para 25 usuarios Con SQL Standard	1 año	$7884,00

IBM – Cognos

Tabla 3

IBM – Cognos descripción

Versión	Información del producto	Vigencia del contrato	Precio por usuario
IBM Cognos Express Business Intelligence User per Authorized User	Informes, planes de instrumentos y consultas en autoservicio para todos los usuarios (Licencia + SW Subscripción & Soporte)	1 año	$633,00
Cognos Express Performance Management User	Modelado hipotético Información racionalizada de recopilación, agregación y análisis Entorno de planificación intuitivo Completa integración con Microsoft Excel	1 año	$1148,00
IBM Cognos Analytics on Cloud	Mínimo 25 usuarios. Incluye: • Ad hoc reporting • Extensible visualization • Dashboards •User and role management •Mail delivery service •Mobile applications •User storage up to 100GB •Disaster recovery up to 100GB	1 año	$12000,00

Código Abierto (Open Source)

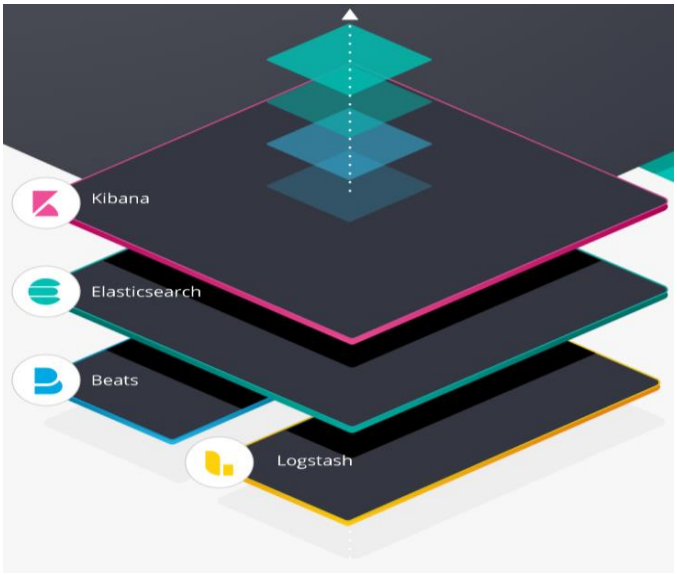
Dentro de las herramientas que serán consideradas para la implementación de la plataforma inteligente y que cumplen con las características de poder ser montadas en la nube y de ser de código abierto, tenemos las siguientes:

ELK Stack (Beats, Logstash, Elasticsearch, Kibana)

Es una colección de software de código abierto que ayuda a proporcionar información en tiempo real sobre los datos que pueden ser estructurados o no estructurados. Uno puede buscar y analizar datos usando sus herramientas con extrema facilidad y eficiencia.

Figura 5

Visión global del Stack ELKB



Nota. (Elastic, Elastic, 2019)

Beats

Se conoce a los Beats como cargadores de datos livianos, está desarrollado en Go, se debe instalar en sus respectivos servidores para capturar todo tipo de datos:

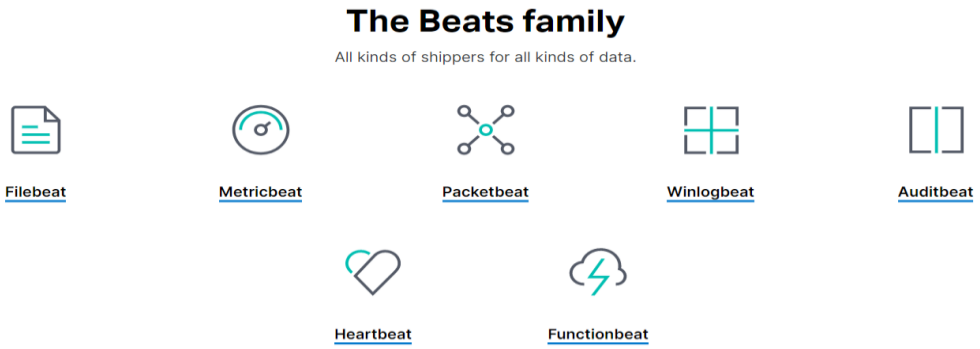
- Registros
- Métricas
- Datos de paquetes de red

Los Beats son capaces de enviar datos directamente a Logstash o Elasticsearch para que a su vez puedan ser visualizados en Kibana.

Se los conoce como livianos porque tienen una pequeña instalación, además utiliza recursos limitados del sistema y no tiene dependencias de tiempo de ejecución. Existen varios tipos de Beats que están oficialmente desarrollados por el equipo de Elasticsearch, pero la comunidad de desarrolla ha ido creado nuevos o una variación de los Beats, a continuación, una lista de los Beats a hoy en día:

Figura 6

Familia de Beats



Nota. *(Elastic, Elastic, 2019)*

Cada Beat tiene una funcionalidad diferente y de gran utilidad a continuación se realiza un pequeño resumen de cada uno:

Tabla 4

Descripción de Beats

Beats	Descripción
Auditbeat	Recopila los datos del marco de auditoría de Linux, Windows y monitorea la integridad de sus archivos.
Filebeat	Lee archivos de registros de colas y logs.
Functionbeat	Lea y envíe eventos desde la infraestructura sin servidor.
Heartbeat	Realiza pings remotos para verificar disponibilidad.
Metricbeat	Obtiene conjuntos de métricas del sistema operativo y los servicios.
Packetbeat	Monitorea la red y los paquetes de las aplicaciones.
Winlogbeat	Obtiene y envía registros de eventos de Windows.

Nota. *(Elastic, Elastic Beats, 2019)*

Elasticsearch

Es un motor de búsqueda y análisis RESTful[7] distribuido, capaz de resolver un número creciente de casos de uso. Como el corazón de Elastic Stack, almacena centralmente sus datos para que pueda descubrir lo esperado y encubrir lo inesperado. Elasticsearch le permite realizar y combinar muchos tipos de búsquedas: estructuradas, no estructuradas, geográficas, métricas, etc. Se basa en el lenguaje de programación Java, lo que permite que Elasticsearch se

[7] RESTful: define un conjunto de principios arquitectónicos por los que se pueden diseñar servicios Web

ejecute en diferentes plataformas. Permite a los usuarios explorar una gran cantidad de datos a muy alta velocidad.

Logstash

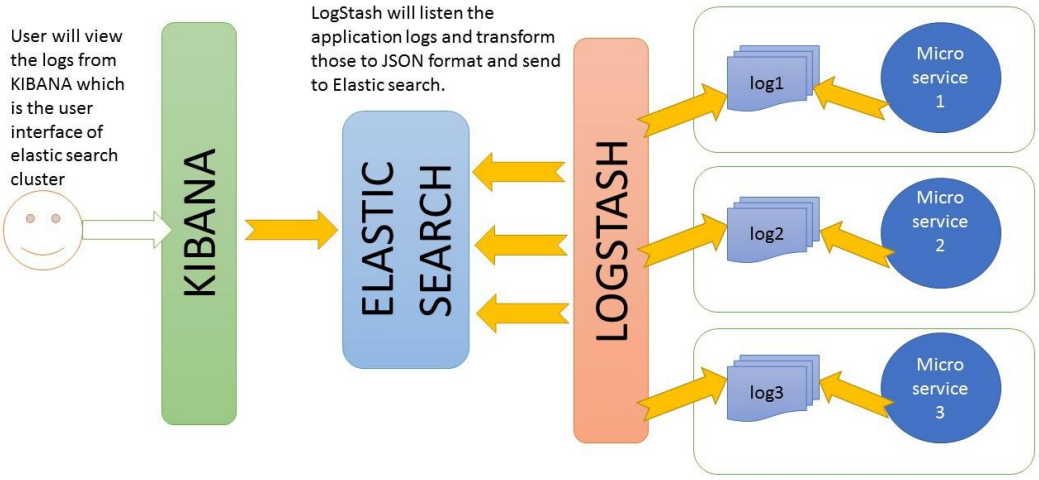
Es una fuente de procesamiento de datos de código abierto del lado del servidor que ingiere datos de una multitud de fuentes simultáneamente, la transforma y luego la envía a su "repositorio" favorito (como Elasticsearch). Los datos a menudo se encuentran dispersos o almacenados en muchos sistemas y en muchos formatos. Logstash admite una variedad de entradas que extraen eventos de una multitud de fuentes comunes, todas al mismo tiempo. Ingiere fácilmente de sus registros, métricas, aplicaciones web, almacenes de datos y diversos servicios de AWS, todo en forma continua y en tiempo real. Logstash tiene un marco conectable con más de 200 complementos. Permite mezclar, combinar y organizar diferentes entradas, filtros y salidas para trabajar en armonía de pipelines.

Kibana

Es una plataforma de análisis y visualización de código abierto diseñada para trabajar con Elasticsearch. Kibana se utiliza para buscar, ver e interactuar con los datos almacenados en los índices de Elasticsearch. Puede realizar fácilmente análisis de datos avanzados y visualizar sus datos en una variedad de cuadros, tablas y mapas. Kibana facilita la comprensión de grandes volúmenes de datos. Su sencilla interfaz basada en navegador le permite crear y compartir rápidamente cuadros de mando dinámicos que muestran cambios en las consultas de Elasticsearch en tiempo real (Gormley & Tong, 2015).

Figura 7

ELK interacción con diferentes archivos log de aplicaciones.



ELK stack interaction with different applications based on Log file

Amazon Elasticsearch Service

Amazon Elasticsearch Service es un servicio completamente administrado que le facilita la implementación, la seguridad y la operación de Elasticsearch a escala sin tiempo de inactividad. El servicio ofrece las API de Elasticsearch de código abierto, el complemento Kibana administrado e integraciones con Logstash y otros servicios de AWS, lo que le permite incorporar datos de forma segura de cualquier fuente y buscarlos, analizarlos y visualizarlos en tiempo real. Amazon Elasticsearch Service le permite pagar solo por lo que usa, no hay costos iniciales ni requisitos de uso. Con Amazon Elasticsearch Service, obtiene la pila de ELK que necesita, sin la sobrecarga operativa (AWS, Amazon Web Services, 2019).

Dentro de los principales beneficios que Amazon Elasticsearch Service ofrece, se encuentran los siguientes:

- Implementación y administración sencilla.

- Integración con herramientas de código abierto y servicios de AWS.
- Fácilmente escalable.
- Seguridad y conformidad.
- Alta disponibilidad.
- Rentabilidad.

Funcionamiento

Figura 8

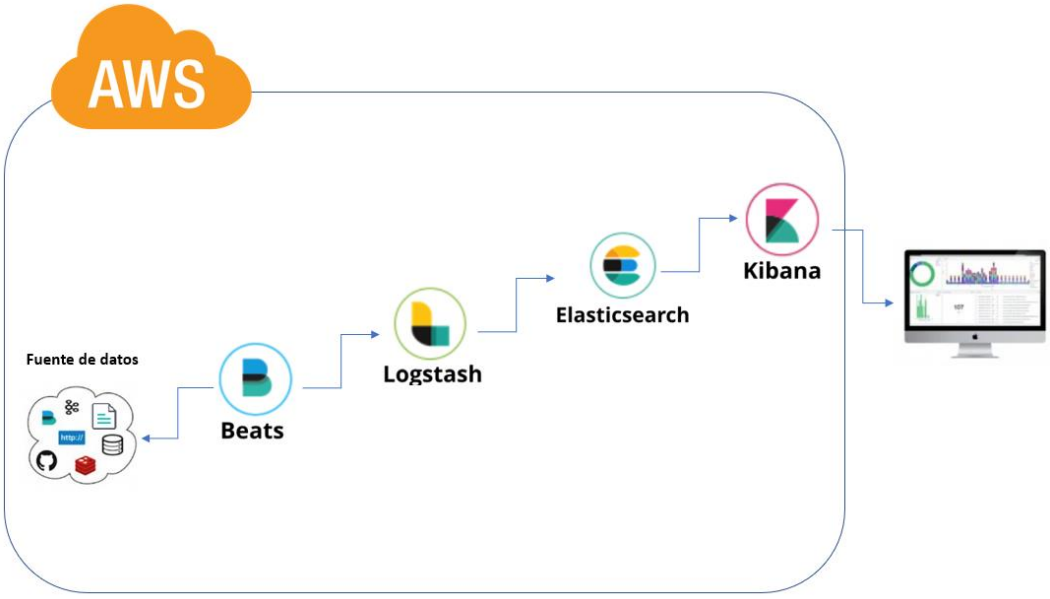
Esquema de funcionamiento de Amazon ElasticSearch Service

Capítulo 3 – Diseño del prototipo

Diseño de la arquitectura basada en la nube

Figura 9

Diagrama de arquitectura



Fuentes de datos

La fuente de los datos puede ser muy diversa ya sean datos estructurados o no estructurados. Estos datos estarán alojados en servidores, repositorios, bases de datos, logs etc. En este punto los datos no poseen ningún valor, para ello Beats nos será de gran ayuda, este módulo leerá los datos y podrá darles un primer tratamiento a los datos.

Para el presente proyecto, las fuentes de datos no están ligadas a un PYME específica, por tal motivo la fuente de datos puede pertenecer muy diversa de acuerdo con cada giro del negocio de cada PYME, tan solo bastará con los lineamientos que uno de los expertos de la empresa nos provea para poder realizar el tratamiento a los datos.

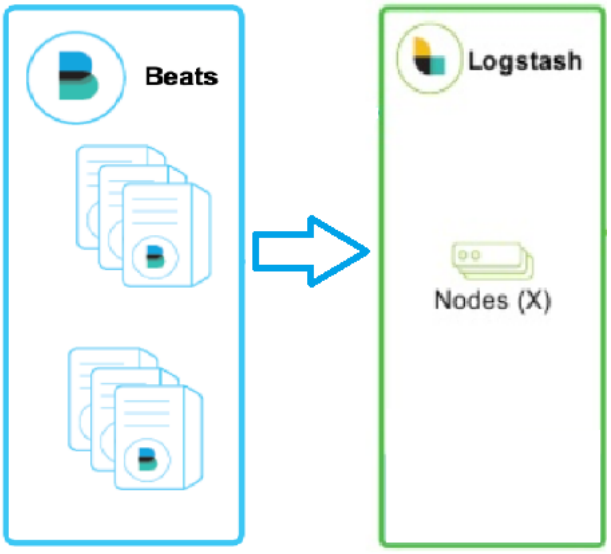
Beats

Beats es de gran importancia ya que posee varias características y tipos de Beats que se pueden usar y con ello solventar algún tipo de necesidad que el cliente solicite.

Como se mencionó anteriormente Beats será el encargado de leer los datos y realizar un pequeño tratamiento a los mismo, tal como ignorar ciertos registros, aplicar expresiones regulares, cabe recalcar en aquí no se debe realizar el tratamiento a los datos ya que el encargado de realizar esta tarea es Logstash.

Figura 10

Paso de datos de Beats a Logstash



Para este proyecto usaremos filebeat ya que es el Beat que nos brinda las características que necesitamos para tratar a los datos de diversas PYMEs este beat es ampliamente usado en el mundo por sus fortalezas y bondades.

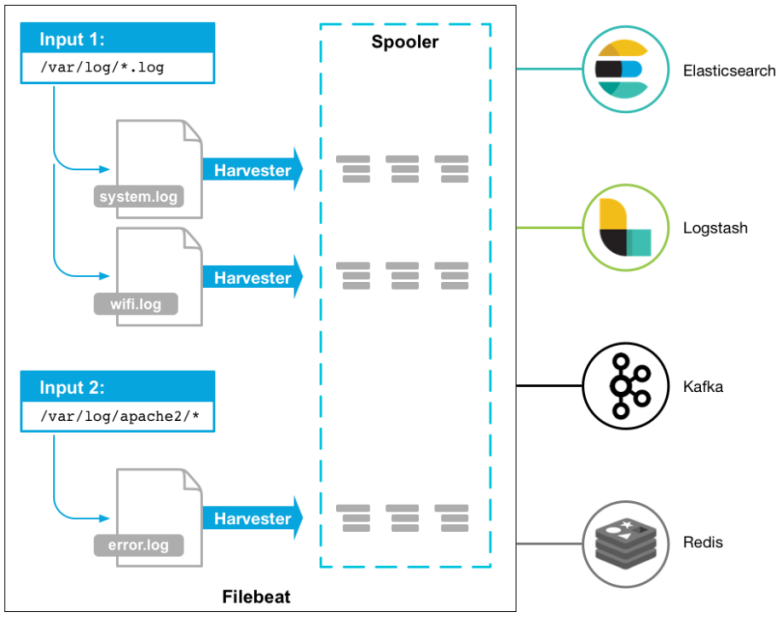
Filebeat

Es un ligero cargador que se encarga de centralizar y reenviar los datos de registro. Se lo instala como un agente en los servidores, también es capaz de supervisar los archivos de registro o las ubicaciones que nosotros especifiquemos, una vez que los recopila los puede reenvía a Elasticsearch o Logstash para su respectiva indexación.

Así es como funciona Filebeat: cuando inicia Filebeat, inicia una o más entradas que se ven en las ubicaciones que ha especificado para los datos de registro. Para cada registro que Filebeat localiza, Filebeat inicia una cosechadora. Cada recolector lee un único registro para contenido nuevo y envía los nuevos datos de registro a libbeat, que agrega los eventos y envía los datos agregados a la salida que ha configurado para Filebeat (Elastic Docs, 2019).

Figura 11

Como trabaja Filebeat



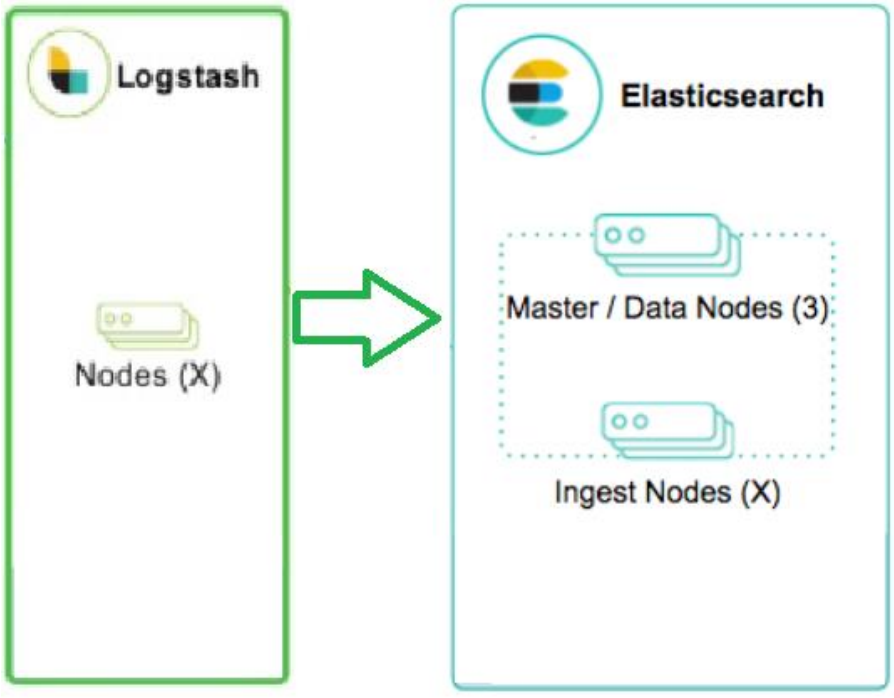
Nota. (Elastic Docs, 2019)

Logstash

Logstash recibe como entrada los datos que son proveídos por Beats y será el encargado de realizar el tratamiento de los datos a profundidad, se encargará de depurar la data según sea la necesidad del cliente, para que posteriormente se almacene en Elasticsearch. Se podría leer directamente los datos desde Logstash sin pasar por Beats, pero se perdería todas las bondades que Beats proporciona.

Figura 12

Paso de datos de Logstash a Elasticsearch



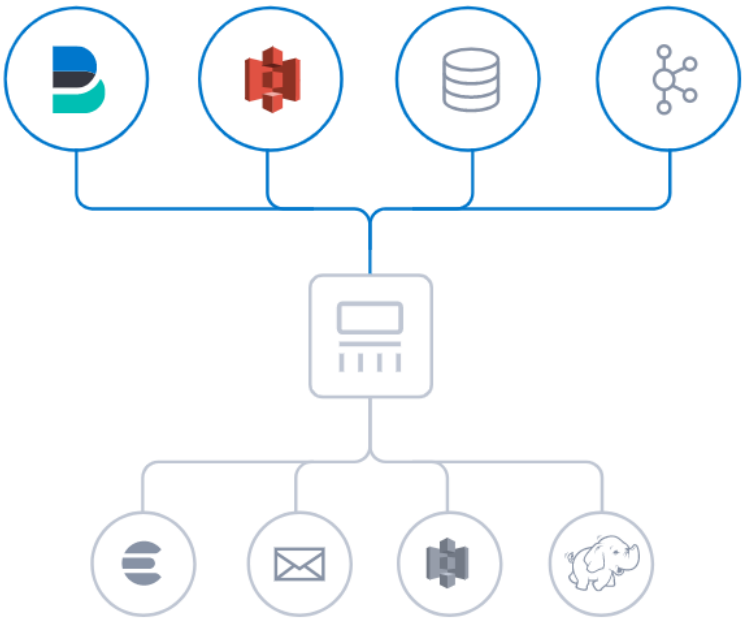
Para este proyecto no se van a establecer filtros previamente definidos, los patrones y filtros que se vaya a aplicar, dependerá de la decisión y necesidades de la PYME, un especialista de la empresa será en el cargado de definir estos filtros y patrones. Se debe definir un especialista por cada empresa.

Input

Logstash tiene la capacidad de procesar al mismo tiempo diferentes datos de entrada como bases de datos, archivos de logs, repositorios, archivos de texto, etc.

Figura 13

Diferentes inputs soportados por Lostash



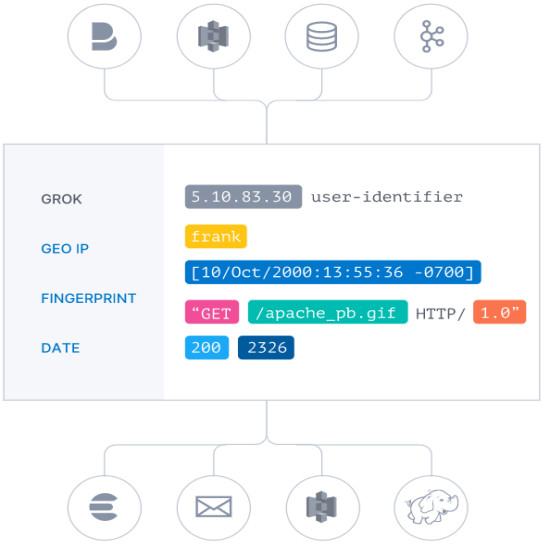
Nota. (Elastic, Elastic, 2019)

Filter

En este apartado se realiza el parseo y transformación dinámica e independientemente de su estructura o complejidad. Se aplican patrones o expresiones regulares de acuerdo con lo que solicite el especialista de la PYME, mediante grok podemos aplicar patrones o expresiones regulares, también Logstash puede descifrar las coordenadas geográficas mediante IP.

Figura 14

Vista de Filter dentro de Logstash



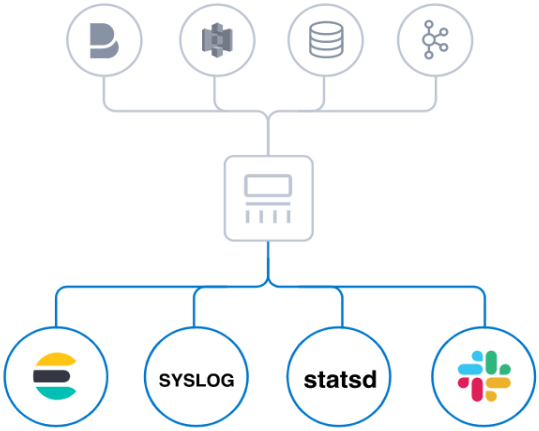
Nota. (Elastic, Elastic, 2019)

Output

Logstash es capaz de soportar una gran variedad de salidas gracias a sus diversos plugins como por ejemplo archivos csv, Elasticsearch, email, http, Kafka, consola, websocket, etc.

Figura 15

Diferentes Outputs soportados por Logstash



Nota. (Elastic, Elastic, 2019)

Elasticsearch

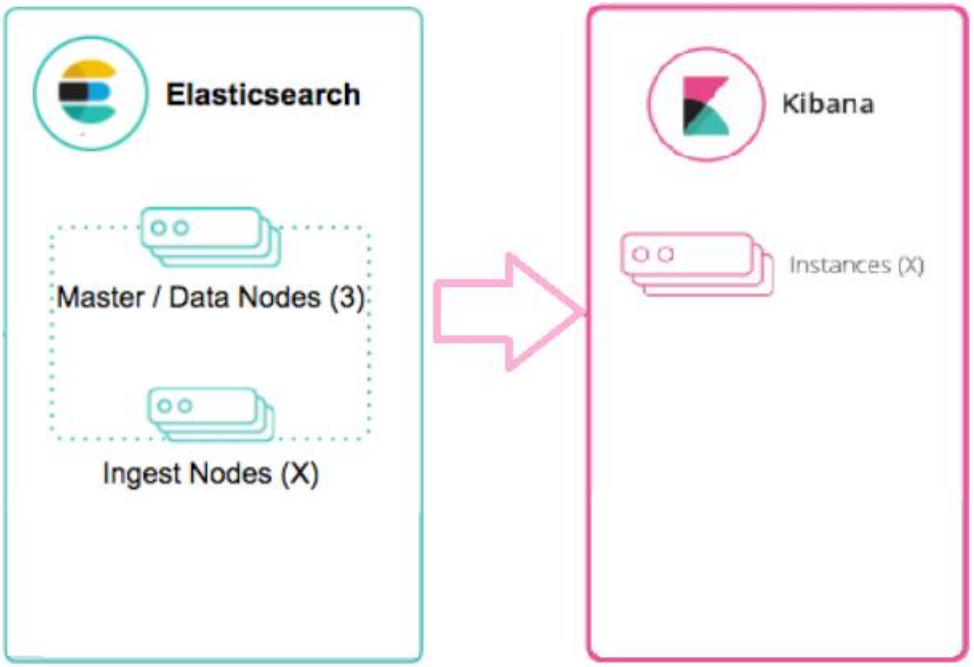
En Elasticsearch será donde alojemos nuestros datos que previamente fueron recolectados y tratados, es una base de datos distribuida, se encarga de distribuir toda la información en cada uno de los nodos por lo tanto es:

- Tolerante a fallos
- Alta disponibilidad

Por otro lado, no solo distribuye la información si no también distribuye el procesamiento. Al realizar una consulta o búsqueda si la información se encuentra distribuida, cada uno de los nodos será el que procese la información y retorne los resultados, teniendo mejores rendimientos.

Figura 16

Paso de datos de Elasticsearch a Kibana



Kibana

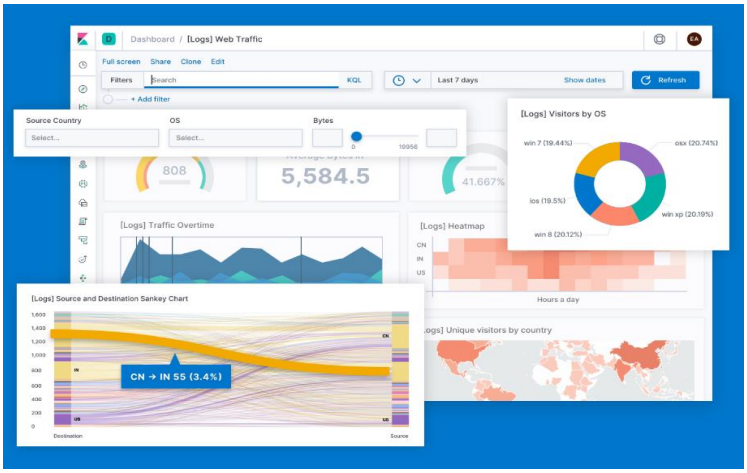
Los resultados se visualizarán en Kibana en un dashboard dinámico, en el cual se podrá realizar filtros e histogramas para una mayor comprensión de los datos que fueron analizados, será de gran ayuda en la toma de decisiones a nivel operacional y gerencial.

Además, es un sistema analítico de código abierto que permite hacer análisis y búsquedas creando dashboard dinámicamente. Se considera Kibana como una plataforma de análisis y visualización de código abierto que trabaja con Elasticsearch. Tiene la capacidad de realizar análisis de datos avanzados y visualizar los datos en una variedad de gráficos, tablas y mapas.

Es de gran ayuda porque nos facilita el análisis de grandes volúmenes de datos gracias a su interfaz simple permite crear y compartir rápidamente paneles dinámicos que muestran cambios en las consultas de Elasticsearch en tiempo real (Elastic Docs, 2019).

Figura 17

Visualizaciones en Kibana



Nota. (Elastic, Elastic, 2019)

Diseño de modelos de datos

Como el presente proyecto trata de una plataforma orientada a la nube para brindar soluciones de Inteligencia de Negocios a PYMES, definir un único modelo de datos resulta inapropiado tomando en cuenta que para cada PYME tanto las fuentes de datos (Bases de datos, archivos planos, APIs, etc) como las reglas de negocio son diferentes; es necesario definir un modelo genérico que pueda aplicarse particularmente a cada PYME.

Modelo de datos

Considerando lo anterior, hemos utilizado la metodología CRISP-DM (Cross Industry Standard Process for Data Mining). Porque proporciona una descripción del ciclo de vida de un proyecto de análisis de datos de forma normalizada, además contempla el proceso de análisis de datos como un proyecto profesional, estableciendo así un contexto mucho más rico que influye en la elaboración de los modelos de datos.

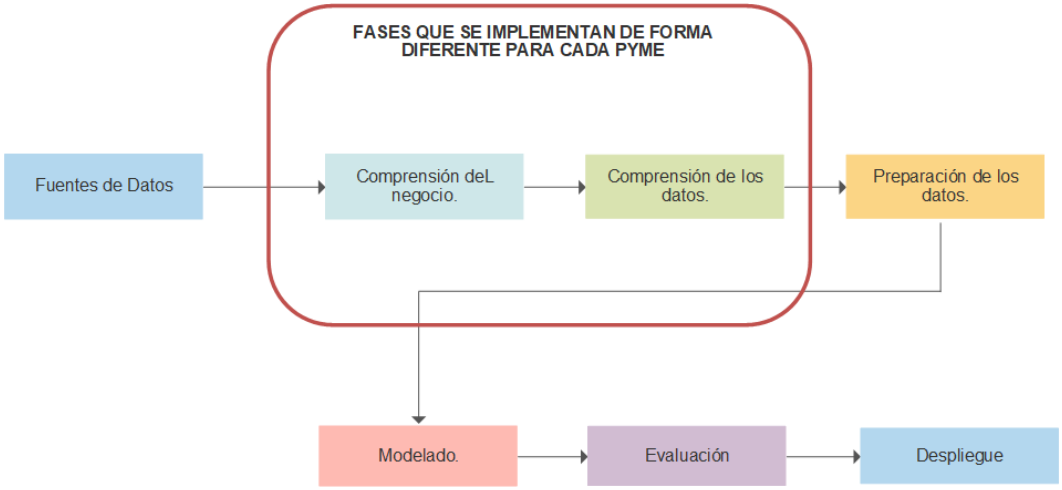
CRISP-DM define las siguientes fases dentro del ciclo de un proyecto:

- Comprensión de los datos (Business understanding).
- Comprensión del negocio (Data understanding).
- Preparación de los datos (Data preparation).
- Modelado (Modeling).
- Evaluación (Evaluation).
- Despliegue (Deployment).

En base a esto el modelo de datos que se definió para el proyecto queda de la siguiente manera:

Figura 18

Modelo de datos



En dónde las fases 1 y 2, cambian o se implementan de forma particular según las fuentes de datos y giro de negocio de la PYME.

Metodología

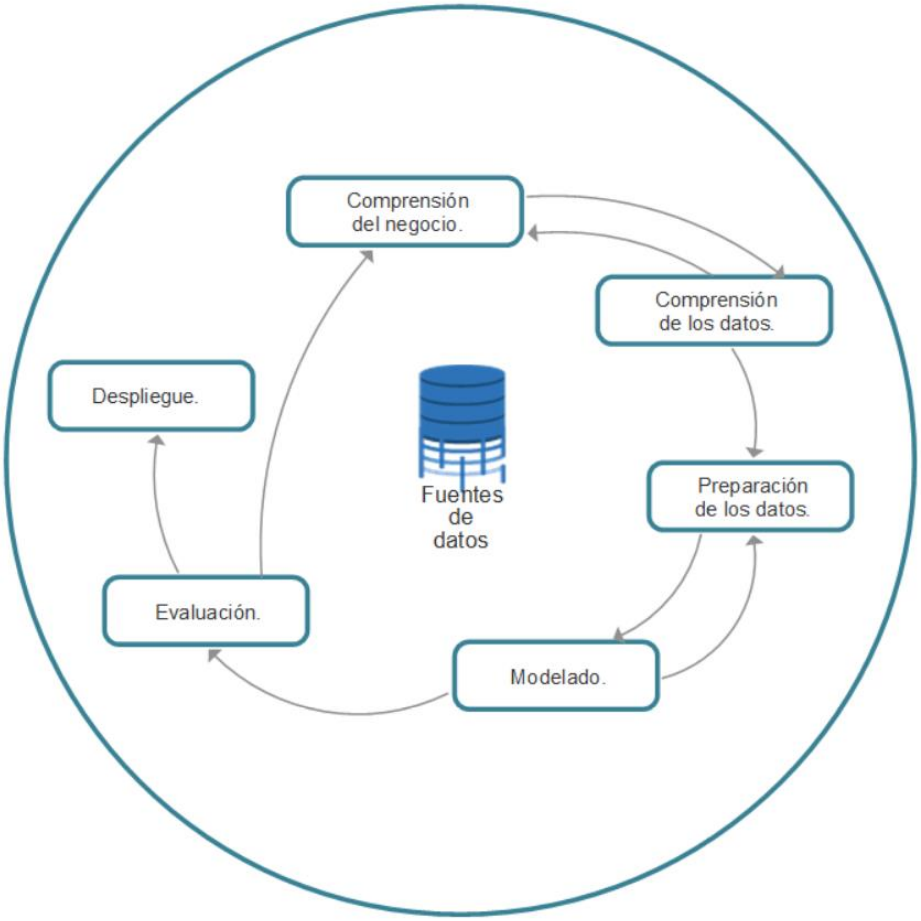
Definición

La metodología CRISP-DM (Cross Industry Standard Process for Data Mining) contempla el proceso de análisis de datos como un proyecto profesional, estableciendo así un contexto mucho más rico que influye en la elaboración de los modelos. Este contexto tiene en cuenta la existencia de un cliente que no es parte del equipo de desarrollo, así como el hecho de que el proyecto no sólo no acaba una vez se halla el modelo idóneo ya que después se requiere un despliegue y un mantenimiento (Azevedo & Santos, 2008).

Esta metodología describe las fases para el proyecto, así como también las tareas necesarias para cada fase y explica la relación entre estas.

Figura 19

Ciclo vital del modelo



El ciclo de vida de este modelo está compuesto de seis fases con flechas que indican las relaciones y dependencias más frecuentes entre ellas. El orden de ejecución o secuencia de las fases no es necesariamente estricto. Lo que permite a la mayoría de los proyectos avanzar o retroceder entre fases si así lo consideran necesario. A continuación, se describen cada una de las fases en que se divide CRISP-DM.

Fase de comprensión del negocio o problema

La primera fase de la guía de referencia CRISP-DM, denominada fase de comprensión del negocio o problema (figura 2.6), es presumiblemente una de las más importantes fases e incluye las tareas de entendimiento de los objetivos y requerimientos del proyecto desde un punto de vista empresarial o institucional, con la finalidad de transformarlos en objetivos técnicos y en un plan de proyecto. Si no se consigue comprender dichos objetivos, será complicado obtener resultados de los cuales fiarse.

Para obtener el mejor provecho de Data Mining, es necesario entender de manera profunda el problema al que se va a dar solución, esto permite acopiar los datos adecuados y realizar una correcta interpretación de los resultados.

Dentro de esta fase es de vital importancia poder transformar el conocimiento obtenido del negocio, en un problema de Data Mining y en un plan previo cuyo objetivo sea el alcanzar las metas del negocio (Gallardo Arancibia, 2009).

Una breve descripción de las principales tareas que intervienen en esta fase es la siguiente:

- *Determinar los objetivos del negocio.* Es la primera tarea a realizar, cuyo objetivo es determinar el problema que se va a resolver.

- *Evaluación de la situación*. En esta tarea se debe graduar el estado de la situación actual, considerando: el conocimiento previo disponible acerca del problema, la cantidad de datos necesaria para dar solución al

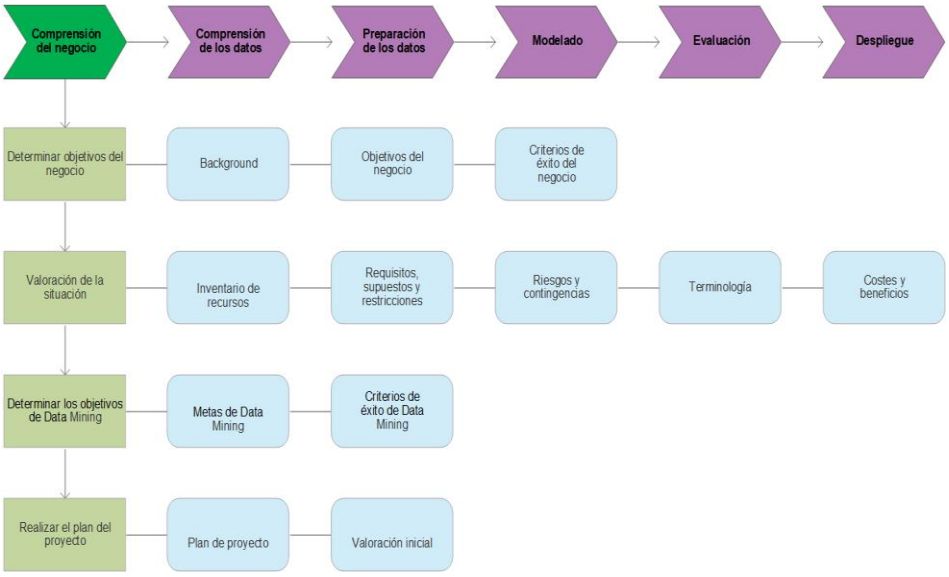
problema, etc. En esta fase se definen los requisitos del problema, tanto en términos de negocio como en términos de Data Mining.

- *Determinación de los objetivos de Data Mining.* La principal meta de esta tarea es la de representar los objetivos del negocio en cuestiones de las metas del proyecto de Data Mining.

- *Producción de un plan del proyecto.* Finalmente, esta última tarea de la primera fase de CRISP-DM, tiene como meta desarrollar un plan para el proyecto, que describa los pasos a seguir y las técnicas a emplear en cada paso.

Figura 20

Comprensión del negocio y sus tareas



Fase de comprensión de los datos

Abarca la recopilación preliminar de los datos, con el propósito de entablar un contacto inicial con el problema, adaptándose con ellos, identificando su calidad y estableciendo las relaciones más obvias que ayuden a determinar las primeras hipótesis (Gallardo Arancibia, 2009).

Las principales tareas de esta fase son:

- *Recolección de datos iniciales.* Esta tarea se enfoca en la recolección inicial de los datos y su acondicionamiento para su posterior procesamiento. Tiene como objetivo, crear informes con un listado de los datos obtenidos, su ubicación, las estrategias usadas en su recolección y los inconvenientes y soluciones relacionados a este proceso.

- *Descripción de los datos.* Una vez obtenidos los datos preliminares, estos pasan a ser descritos, lo que involucra crear volúmenes de datos, el significado de cada campo y la explicación del formato inicial.

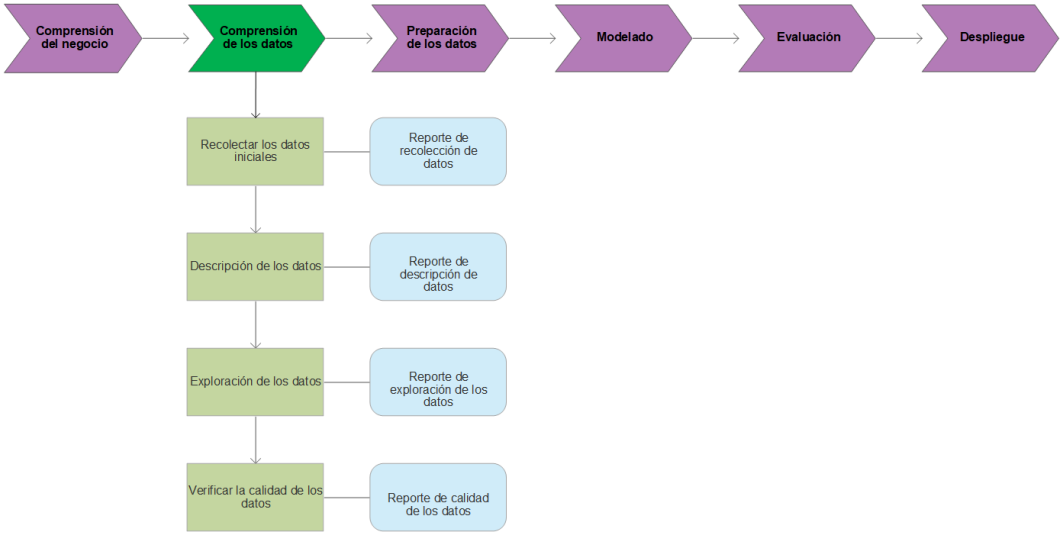
- *Exploración de datos.* La finalidad de esta tarea es determinar una estructura genérica para los datos, para lo que se deben aplicar pruebas que manifiesten las propiedades en los datos adquiridos, se crean gráficos de distribución y se construyen tablas de frecuencia.

- *Verificación de la calidad de los datos.* Se realizan confirmaciones sobre los datos, para establecer la coherencia de los valores de cada campo, la cantidad de valores nulos y su distribución, y para encontrar valores fuera

de rango, que pueden convertirse en información no fiable, la principal meta de esta tarea es asegurar la consistencia y completitud de los datos.

Figura 21

Comprensión de los datos y sus tareas



Fase de preparación de los datos

Una vez realizada la recolección de datos, continúa su preparación para adecuarlos a las técnicas de Data Mining que se utilicen en el futuro, tales como: visualización de datos, de búsqueda de vínculos entre variables u otras medidas para exploración de los datos.

La preparación de datos abarca las actividades de selección de datos para ser aplicados en un determinado proceso de modelado, limpieza de datos, integración de diferentes fuentes y cambios de formato (Gallardo Arancibia, 2009).

Las principales tareas de esta fase son:

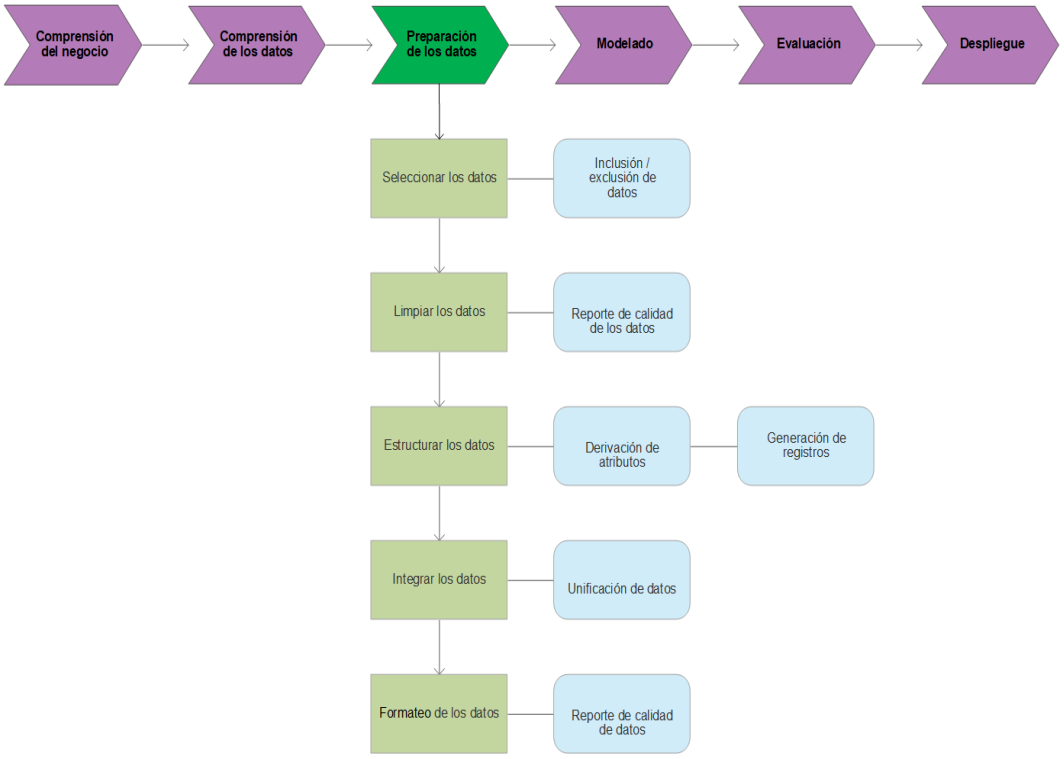
- *Selección de datos.* Esta etapa consiste en seleccionar un subconjunto de los datos preliminares, en base a los siguientes criterios: calidad de los

datos (completitud, corrección y limitaciones de volumen) o en los tipos de datos que están enlazados con las técnicas de Data Mining establecidas.

- *Limpieza de los datos.* Esta tarea es una de las que requiere mayor tiempo y esfuerzo, debido a la gran cantidad de técnicas que se pueden utilizar para mejorar la calidad de los datos con el fin de arreglarlos para la fase de modelado. Dentro de las técnicas que se pueden utilizar para lograr este acometido están: normalización de los datos, discretización de campos numéricos, tratamiento de valores ausentes, reducción del volumen de datos, etc.
- *Estructuración de los datos.* Incluye las técnicas y actividades de preparación de datos, como creación de nuevos atributos a partir de otros existentes, incorporación de nuevos registros o conversión de atributos existentes.
- *Integración de los datos.* Esta tarea se encarga de crear nuevas estructuras basándose en los datos previamente seleccionados, por ejemplo, creación de nuevos campos tomando como referencia otros ya existentes, creación de nuevos registros, unificación de tablas que engloben distintos atributos para un mismo objeto, agregación de nuevas tablas donde se resuman características de varios registros.
- *Formateo de los datos.* La tarea se enfoca principalmente en la transformación de la sintaxis de los datos sin alterar su significado o valor, esto, con la finalidad de posibilitar el uso de alguna técnica de Data Mining en particular, por ejemplo: la reorganización de los registros y/o campos de las tablas o el acoplamiento de los valores según las limitaciones del modelo (borrar comas, espacios en blanco, caracteres especiales, etc.).

Figura 22

Preparación de los datos y sus tareas



Fase de modelado

En esta fase se eligen las técnicas de modelado adecuadas para el proyecto de Data Mining. Las técnicas que se van a emplear en esta fase son seleccionadas basándose en las siguientes pautas:

- Debe ser apropiada al problema.
- Se debe disponer de los datos necesarios.
- Debe adaptarse a los requisitos del problema.
- Ofrecer un tiempo prudente para obtener un modelo.
- Se debe poseer el conocimiento justo y necesario para aplicar la técnica.

Las principales tareas de esta fase son:

- *Selección de la técnica de modelado.* La tarea está orientada a seleccionar una técnica de Data Mining que se acople al problema que se desea resolver. Para lograrlo hay que considerar el objetivo del proyecto y su relación con las herramientas seleccionadas.

- *Generación del plan de prueba.* Con el modelo construido, se generan procesos orientados a probar la calidad y funcionamiento del mismo. Normalmente se crean dos conjuntos de datos, uno para entrenamiento y otro para pruebas, de esta forma se consigue medir el modelo generado con el conjunto de prueba, comparándolo con el modelo basado en el conjunto de entrenamiento.

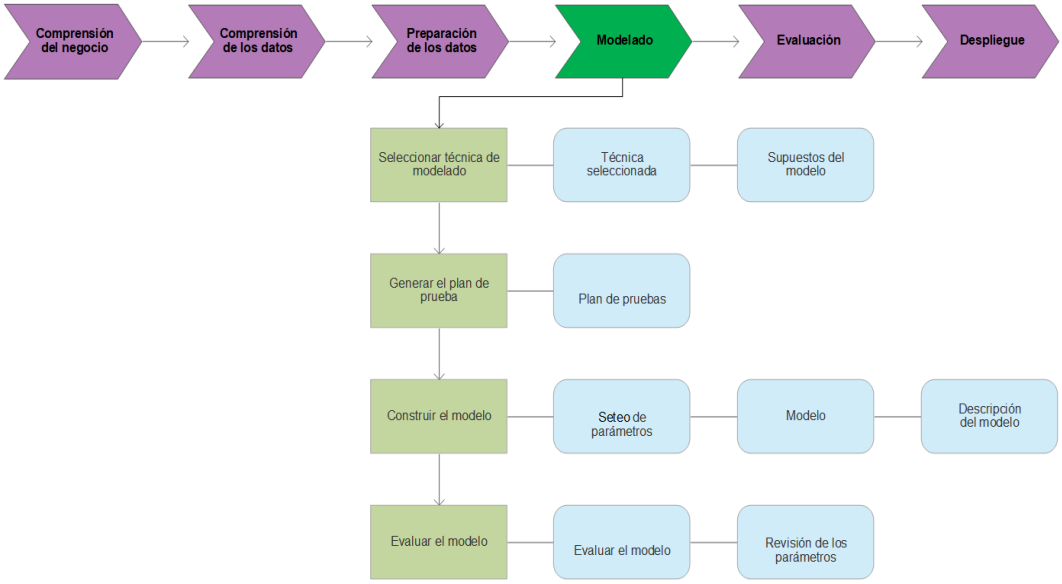
- *Construcción del Modelo*. Una vez obtenida la técnica, se procede a ejecutarla sobre los datos preparados anteriormente, y de esta forma obtener uno o más modelos. Las técnicas para modelar poseen un grupo de parámetros que definen las propiedades que tendrá el modelo generado. Seleccionar los parámetros representa un proceso basado en los resultados obtenidos, mismo que deben ser interpretados.

- *Evaluación del modelo.* La tarea requiere de la interpretación de los modelos obtenidos usando de referencia el conocimiento del dominio del negocio y los casos de éxito del mismo. Expertos en el dominio del

problema juzgan los modelos dentro del contexto del dominio y expertos en Data Mining aplican sus propios.

Figura 23

Modelado y sus tareas



Fase de evaluación

La fase se enfoca principalmente en evaluar el modelo, considerando que se deben cumplir los objetivos y casos de éxito establecidos en el problema. Se toma en cuenta que la integridad esperada para el modelo es aplicable únicamente para el grupo de datos utilizados en el análisis. Es vital validar el proceso considerando los resultados con el fin de identificar errores cometidos, y poder regresar a algún paso previo (Gallardo Arancibia, 2009).

Las principales tareas de esta fase son:

- *Evaluación de los resultados.* En los pasos de evaluación anteriores, se trataron factores tales como la exactitud y generalidad del modelo

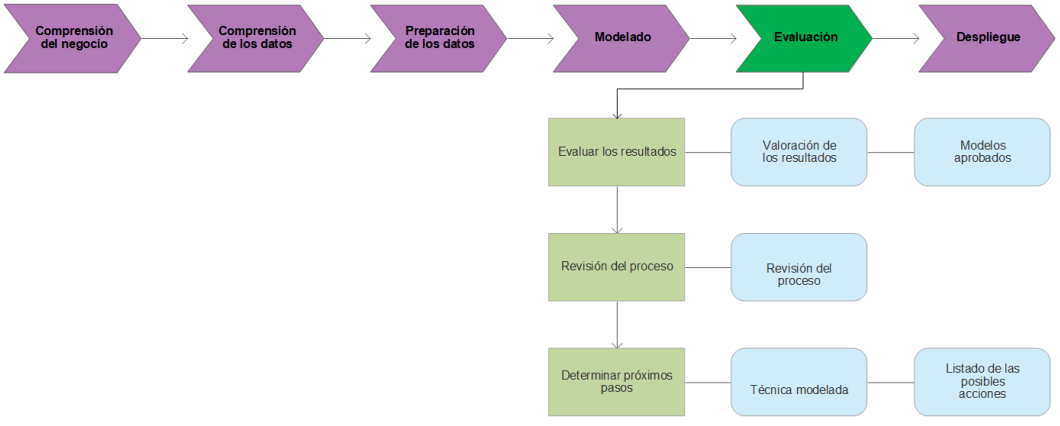
generado, esta tarea se enfoca en la valoración del modelo y las metas del negocio. Además, determina si existe algún criterio por el cual el modelo resulte defectuoso o a su vez identificar si es posible validar el modelo con un ejemplo real si así el tiempo lo amerita.

- *Proceso de revisión.* Dentro de esta tarea se valora el proceso de Data Mining por completo, con la finalidad de identificar factores que puedan someterse a mejoras.

- *Determinación de futuras fases*. Una vez llegada a esta fase, se valoran los resultados obtenidos hasta el momento, con el objetivo de regresar a fases anteriores en caso de obtener resultados no satisfactorios. Si la realidad es la opuesta, se prosigue con la siguiente fase.

Figura 24

Evaluación y sus tareas



Fase de implementación

En esta fase y una vez que el modelo ha sido construido y validado, se convierte el conocimiento y resultados obtenidos en decisiones y acciones en el ámbito del negocio. Estas acciones pueden ser recomendadas por quién analiza los datos, basándose en la apreciación del modelo o bien pueden ser identificadas al aplicar el modelo a diferentes grupos de datos. Como un punto extra dentro de esta fase se debe considerar y asegurar el mantenimiento del proyecto.

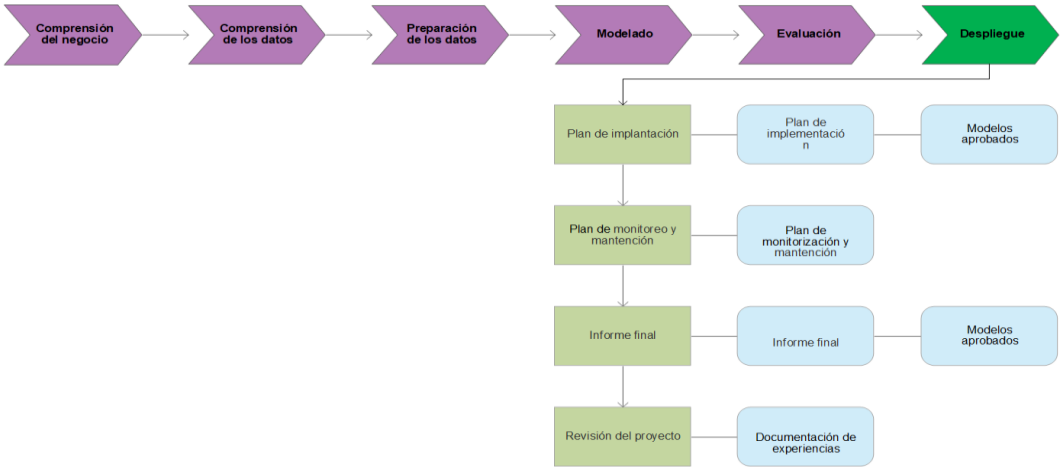
Las principales tareas de esta fase son:

- *Plan de implementación.* Antes de implementar la solución de Data Mining en el negocio, se deben tomar los resultados y finalizar construyendo una estrategia para su implementación. Todos los procesos o procedimientos utilizados para la creación del modelo deben ser documentados para luego ser implementados.
- *Monitorización y Mantenimiento.* Una vez obtenidos los resultados de la solución de Data Mining, después de implementarlos en la organización, es recomendable crear estrategias para monitorear y mantener los modelos. La retroalimentación generada por la monitorización y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente.
- *Informe Final.* Este informe contiene un resumen de las conclusiones y recomendaciones logradas con el proyecto, además de los resultados obtenidos y la experiencia conseguida durante todo el proceso.

- *Revisión del proyecto.* Al finalizar el proyecto, se reflexiona acerca de los logros y errores encontrados durante todo el proceso, además de identificar los puntos en los que se deberían aplicar mejoras.

Figura 25

Implementación y sus tareas



Capítulo 4 – Desarrollo e implementación

Selección de datos a ser procesado

Para el presente proyecto vamos a hacer uso de bases de datos de PYMEs alojadas en internet los que se encuentran en diferentes formatos lo cual no es un limitante para el Stack ELK. Los datos de pruebas son de accidentes de tránsito en Estados Unidos del año 2019.

Desarrollo de procesos de extracción, transformación y carga de datos

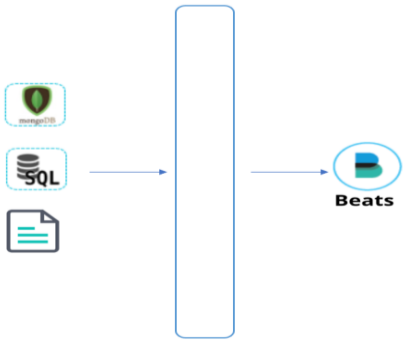
Para el proceso de extracción, transformación y carga de datos se detalla a nivel técnico cuales fueron los respectivos procesos que se usó para obtener buenos resultados de cada uno de ellos.

Extracción de datos

En este proyecto el Beat que usaremos es Filebeat ya que se ajusta a las necesidades que requerimos, será quién nos permita cargar los datos mediante sus diferentes configuraciones por medio de scripts.

Figura 26

Proceso de extracción de datos



Para que Filebeat sea capaz de cargar los datos debemos especificar la ruta del archivo que va a ser cargado, si es de una base de datos se realiza mediante jdbc o cualquiera de los diversos tipos de que soporta filebeat.

Script de configuración para carga de datos con Filebeat

En este script se puede especificar ciertas características que Filebeat puede adoptar a continuación se detalla la configuración adoptada para la carga de datos de datos:

Se debe habilitar a Filebeat para que sea capaz de cargar data, por defecto su estado esta como deshabilitado:

Figura 27

Configuración para habilitar carga de datos

```
# Change to true to enable this input configuration.
enabled: true
```

Se especifica la ruta de donde se va a cargar el origen de los datos como se explicó previamente puede ser di diversos formatos.

Figura 28

Configuración path de origen de datos

```
# Paths that should be crawled and fetched. Glob based paths.
paths:
  - /home/diego/Documentos/ELKB/logs/US_Accidents_May19.csv
  #- c:\programdata\elasticsearch\logs\*
```

Se puede realizar el primer tratamiento o filtro a los datos con el uso de expresiones regulares, estas expresiones regulares se aplicarán de acuerdo a cada PYME, cabe dar a notar que no se debe tratar de hacer un tratamiento

exhaustivo de los datos en Filebeat ya que no es el encargado de esta tarea, quién es el encargado es Logstash.

Figura 29

Configuración de exclusión de primera fila con expresiones regulares

```
# Exclude lines. A list of regular expressions to match. It drops the lines that are
# matching any regular expression from the list.
exclude_lines: ['^ID.*']
```

Definición de la salida (output) de los datos que fueron cargados por Filebeat, aquí se puede definir si los datos van directamente a Elasticsearch o a Logstash, en nuestro caso deberá pasar por primero por Logstash para que posteriormente pase a Elasticsearch.

Figura 30

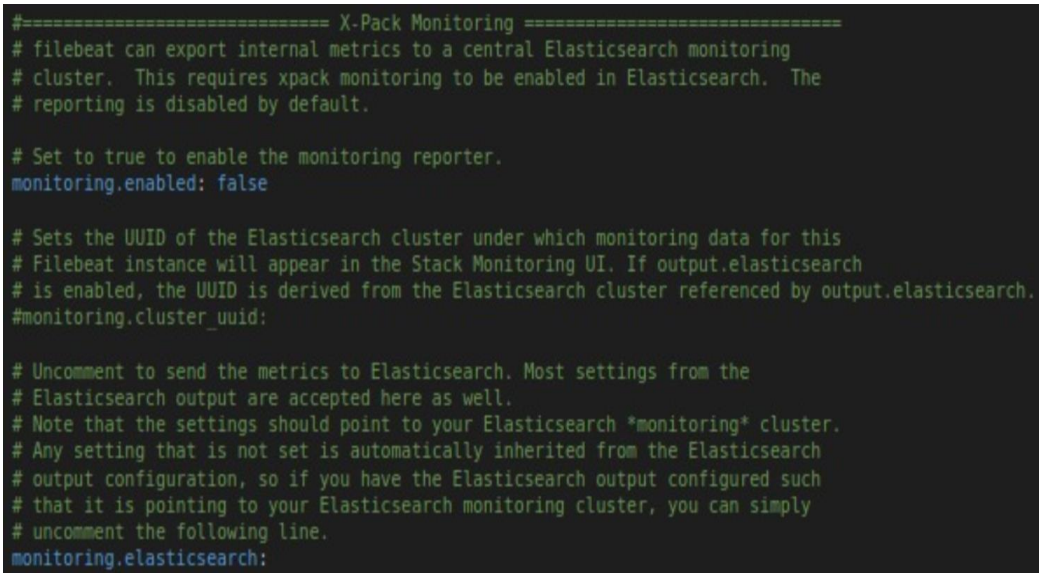
Configuración de la salida (output) de los datos

```
#------------------------------ Logstash output -----------------
output.logstash:
  # The Logstash hosts
  hosts: ["localhost:5044"]

  # Optional SSL. By default is off.
  # List of root certificates for HTTPS server verifications
  #ssl.certificate_authorities: ["/etc/pki/root/ca.pem"]

  # Certificate for SSL client authentication
  #ssl.certificate: "/etc/pki/client/cert.pem"

  # Client Certificate Key
  #ssl.key: "/etc/pki/client/cert.key"
```

Si se pretende poder monitorizar Filebeat desde Kibana se debe realizar una configuración adicional para que se pueda realizar dicha monitorización mediante X-pack.

Figura 31

Configuración monitorización mediante X-pack



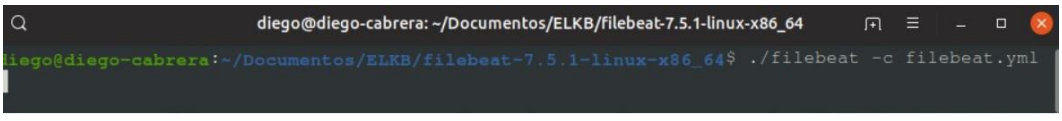
Toda esta configuración es necesaria para poder cargar los datos en Beats y aplicar un pequeño tratamiento a los datos.

Inicio del servicio Filebeat

Mediante la terminal de Linux o cmd en Windows se puede iniciar el servicio de Filebeat mediante el siguiente comando:

Figura 32

Inicio del servicio Filebeat



Transformación de datos

Para que los datos puedan ser transformados de pasar por Logstash en donde se aplican filtros y diversas operaciones que Logstash posee en entre todas

sus funcionalidades, en esta parte nos centramos para tratar los datos y realizar filtros que el especialista de la PYME nos solicite.

Script de configuración para tratar los datos con Logstash

Se debe crear un script en el que va a estar los filtros y tratamiento que se va a dar a los datos, para este proyecto se muestra los filtro y tratamientos que proporcionaron.

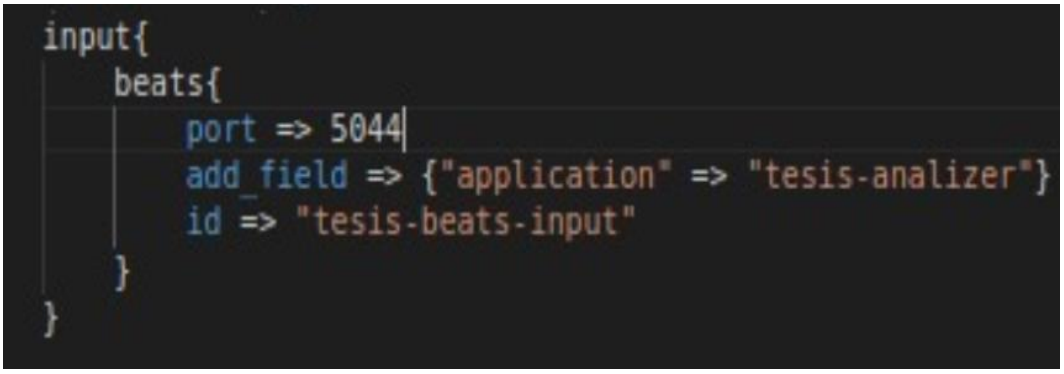
El script se divide en tres secciones las que se detallan a continuación:

Input

Aquí se define de donde provienen los datos de entrada para este proyecto previamente fueron cargados por Filebeat es por ello que ese será nuestro input en Logstash, aquí definiremos también el puerto por el que se levantó Filebeat, se define un Id y también se pueden agregar campos que pueden ser usados en el tratamiento de los datos.

Figura 33

Configuración de input de Logstash



Filter

En esta parte del script será en donde definiremos los patrones y expresiones regulares que ayudaran a realizar el tratamiento de los datos mediante el uso de grok que es uno de los múltiples plugins que posee Logstash,

para poder ir testeando que nuestro patrón o expresión regular aplique a nuestra data podemos usar grok debbuger que es una herramienta que posee Kibana.

Figura 34

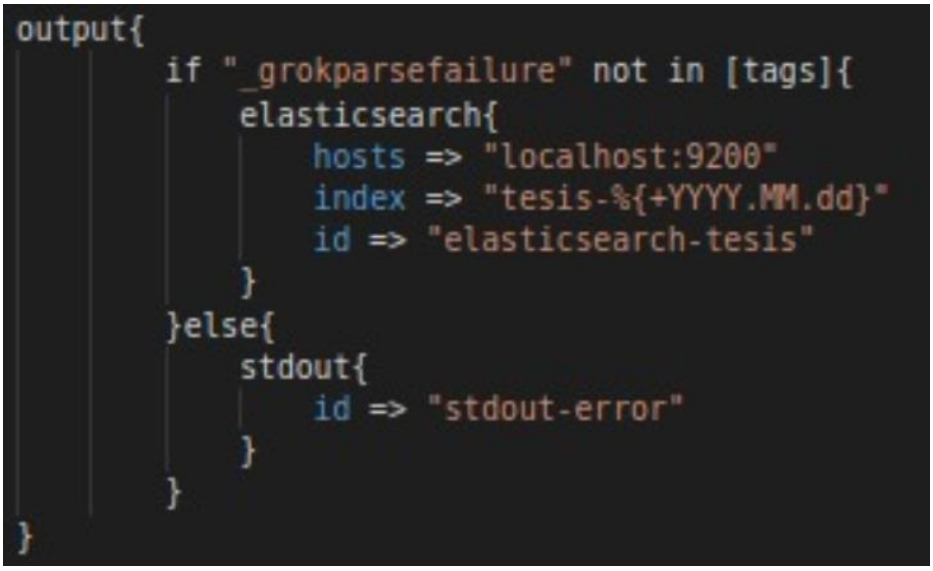
Configuración para tratamiento de los datos

Output

En el output podemos definir a donde se dirigen los datos en nuestro caso a Elasticsearch y por consola, las salidas pueden ser diferentes tipos por la versatilidad de Logstash, su configuración es la siguiente.

Figura 35

Configuración output de Logstash

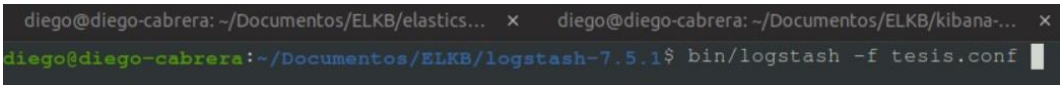


Inicio del servicio Logstash

Para dar inicio al servicio de Logstash se lo realizar mediante el terminal de Linux o cmd en Windows con el siguiente comando:

Figura 36

Inicio del servicio Logstash



Pruebas de transformación de datos

Para realizar pruebas de transformación de los datos haremos uso de Grok Debugger que no permite probar datos de entrada y aplicar los patrones y expresiones regulares y mostrar un resultado en formato JSON si la transformación fue exitosa. A continuación, se muestra un ejemplo en la figura de una data de prueba.

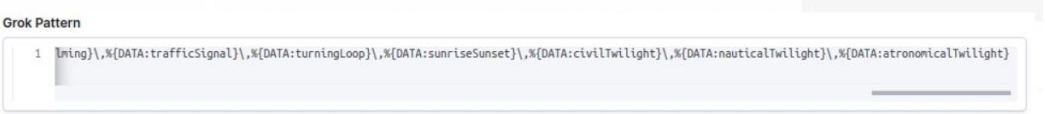
Figura 37

Data de prueba 1

Luego se aplican los patrones y expresiones regulares para el tratamiento de la data.

Figura 38

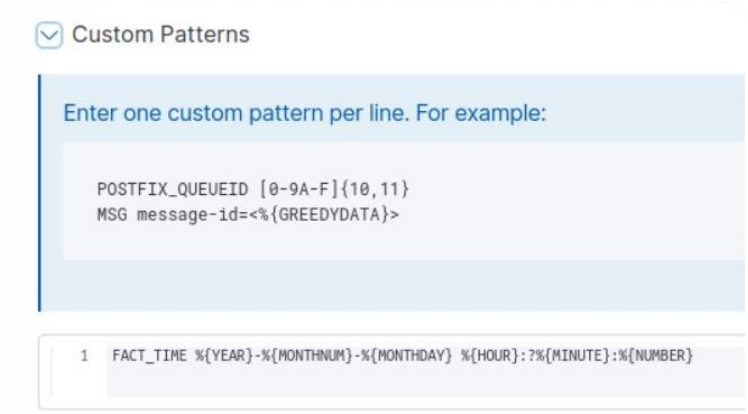
Aplicación de patrones y expresiones regulares



En el caso que los patrones proporcionados por Grog no se adapten a un dato a un requerimiento que solicite el especialista de la PYME podemos personalizar nuestros propios patrones usando expresiones regulares.

Figura 39

Patrón personalizado



Una vez terminado la aplicación de los patrones y expresiones regulares para el tratamiento de la data realizamos la prueba y podremos observar si la transformación de los datos se realizó con éxito.

Figura 40

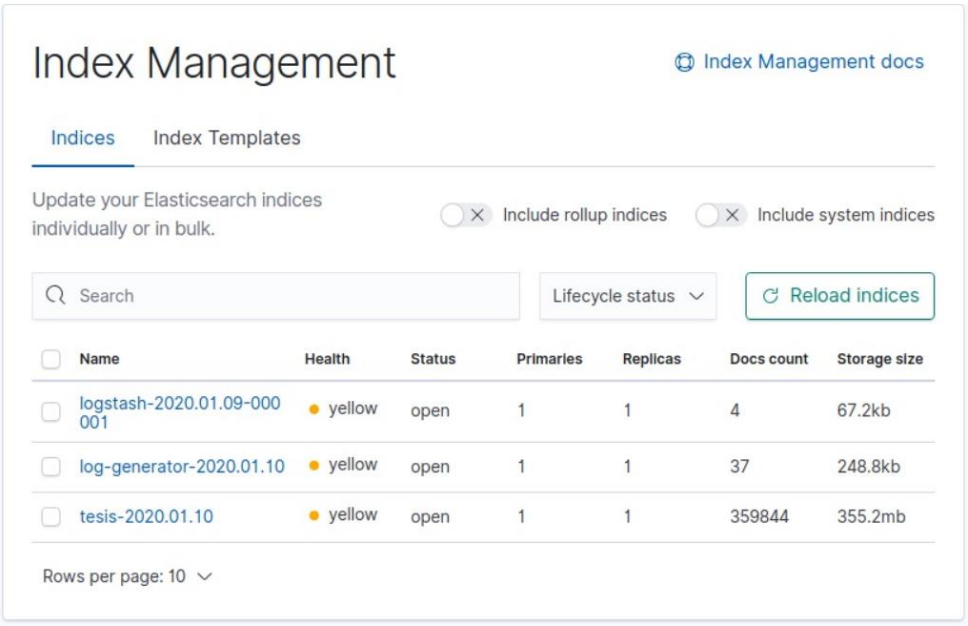
Transformación de datos exitosa

Carga de datos

Para poder hacer una revisión a la carga de datos, lo validaremos mediante la interfaz de administración de Elasticsearch que está alojada en Kibana ahí podemos ver los archivos indexados con el nombre del índice que proporcionamos previamente.

Figura 41

Datos indexados

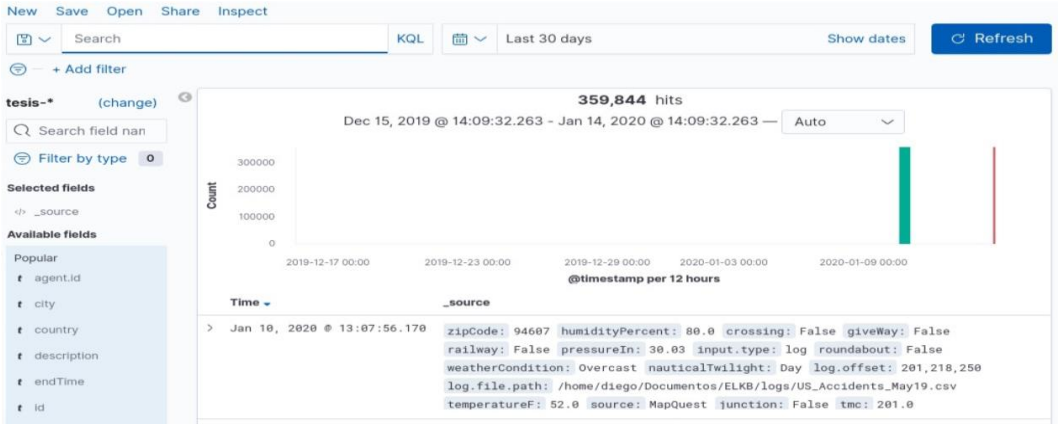


Pruebas de carga de datos

Para poder validar que los datos se cargaron exitosamente se lo hace mediante la interfaz de Kibana la que no proporciona un aspecto visual de los datos que fueron cargados, indexados, tratados y listo para ser visualizados.

Figura 42

Prueba de carga de datos

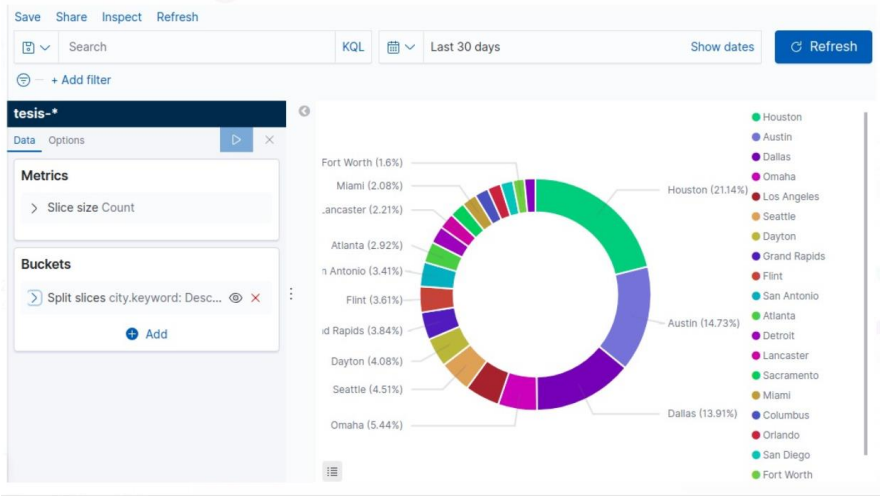


Desarrollo de la plataforma inteligente

Kibana es la encargada de realizar los gráficos de barras, pasteles etc., que el especialista de la PYME nos solicite todo esto se realiza en la interfaz de administración de Kibana ahí gestionaremos los gráficos, visualizaciones y dashboard que nos soliciten, a continuación, se proporciona una muestra de un gráfico y el dashboard que se pueden crear.

Figura 43

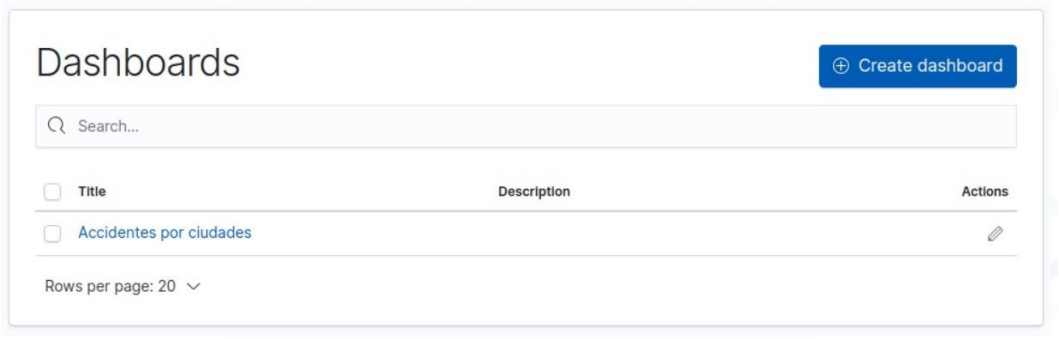
Visualización de gráfico de accidentes por ciudades



Por cada visualización que creamos la podemos ir administrado en un dashborad según las preferencias del cliente.

Figura 44

Creación de dashboard



Implantación de la plataforma en una arquitectura cloud

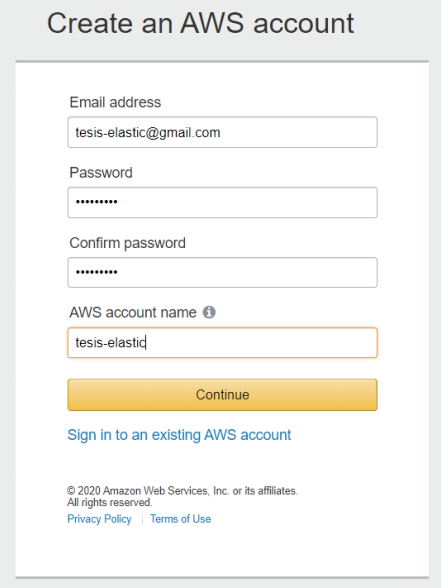
Para la implantación de la plataforma en la nube se decidió usar AWS para ello se tuvo que definir las características a nivel de hardware que se requería, una vez definido este tema a continuación se detalla el procesa de implantación.

Creación de la cuenta en AWS

Para la creación de la cuenta se requiere una tarjeta de crédito internacional para validar la cuenta, en nuestro caso no se pudo utilizar la capa gratis que oferta AWS, porque previamente se definió las características de debía tener nuestra máquina y la capa gratis de AWS no suple nuestras necesidades.

Figura 45

Creación de cuenta e AWS



Luego de crear la cuenta vamos a escoger la imagen de la máquina que vamos a rentar en AWS, para nuestro caso será un de tipo Ubuntu Server 18.04 64 bits

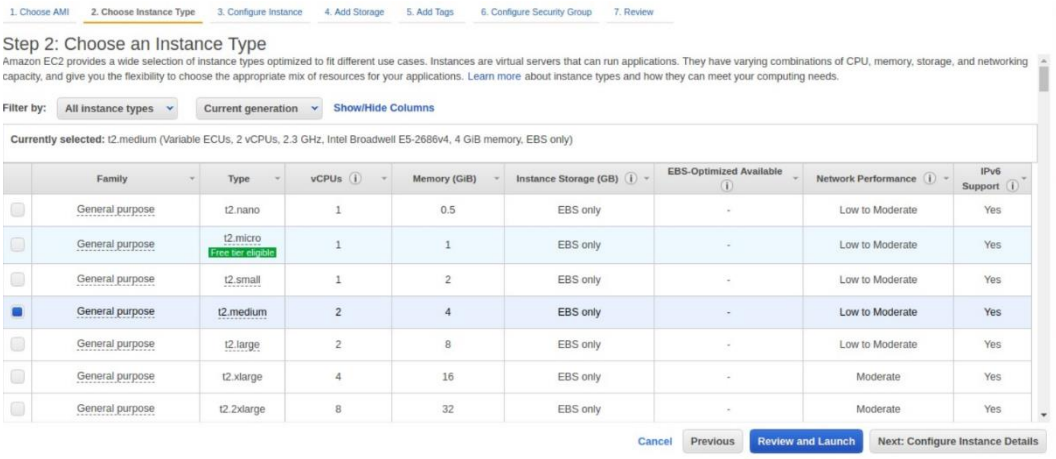
Figura 46

Selección SO máquina virtual

Ya que elegimos el SO de nuestro servidor escogemos el tipo de instancia es decir cuanta RAM vamos a requerir para nuestro proyecto requerimos 4 Gb.

Figura 47

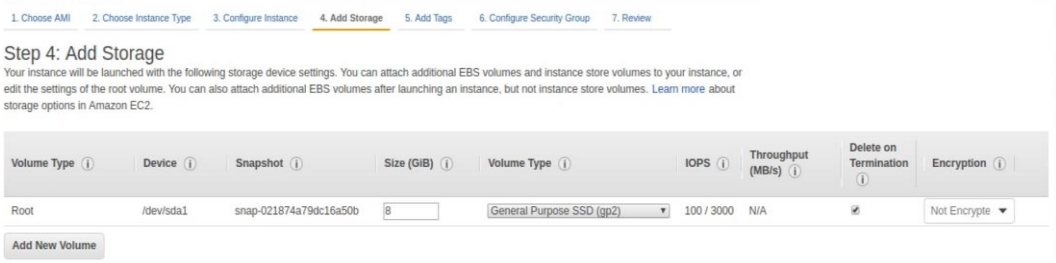
Tipo de instancia y RAM



Ahora tenemos que definir cuanto almacenamiento requerimos, en nuestro caso requerimos un disco SSD con 8 GB.

Figura 48

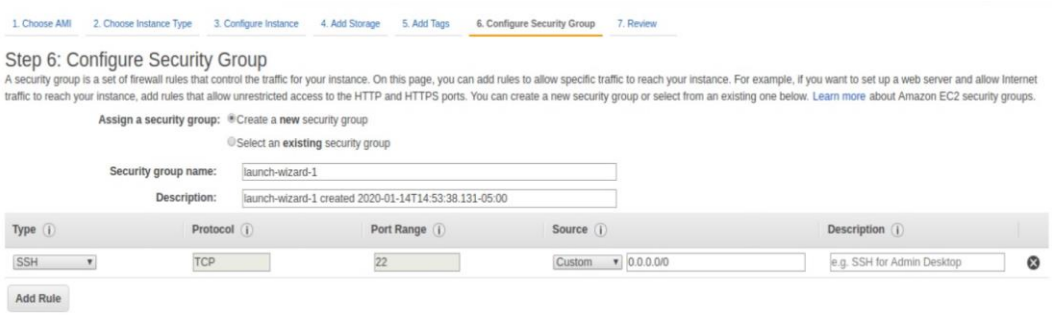
Capacidad y tipo de disco



Y por último se crea una configuración para establecer grupos de seguridad que son de gran importancia para que definir porque protocolos y puertos pueden realizar peticiones a nuestro servidor.

Figura 49

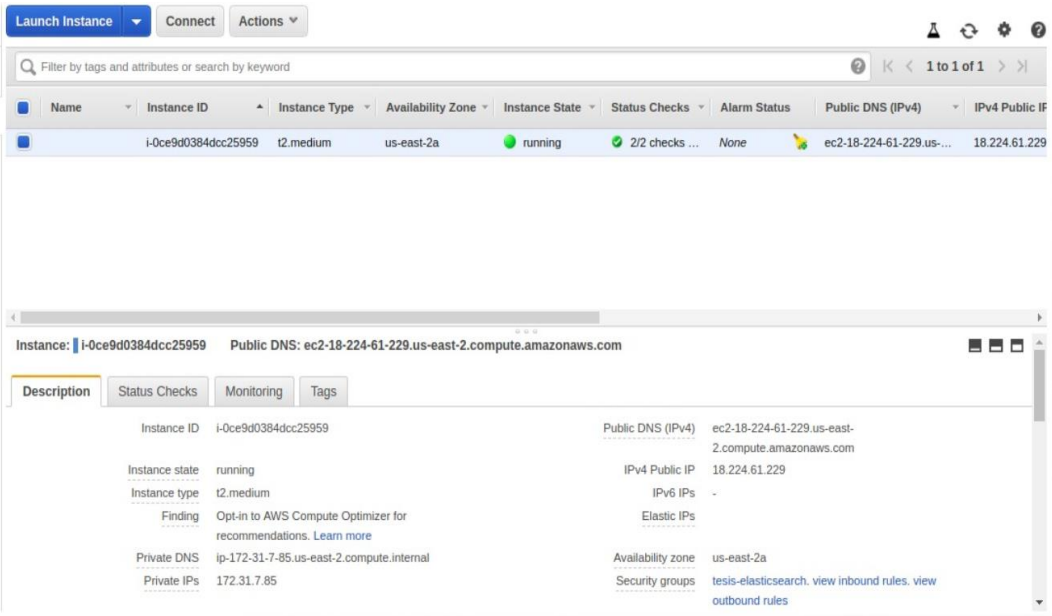
Grupo de seguridad



Una vez que ya se creó nuestra máquina en AWS se nos presentará la siguiente interfaz en la que nos proveen suficiente información para poder conectarnos a nuestra máquina y gestionar diversos aspectos.

Figura 50

Interfaz de administración de servidor en AWS



Discusión de resultados

Una vez terminado este proyecto se pudo determinar que la plataforma propuesta es de gran ayuda para las PYMEs, ya que en su mayoría tienen datos pero solo almacenado en archivos planos, bases de datos etc., sin darles ningún uso adecuado, es por ello que nuestra propuesta genera valor, porque explotamos esos datos para convertirlos en información útil para la empresa, generando valor agregado al analizar información tanto en tiempo real como de datos históricos, con ello las PYMEs podrán tomar decisiones en menores lapsos de tiempo.

El Stack ELK de Elasticsearch nos permite mostrar resultados fáciles de interpretar debido a sus múltiples tipos de visualizaciones que proporcionan información relevante, gráficos de fácil comprensión.

Cada PYME pude realizar un indeterminado número de visualizaciones de acuerdo a cada necesidad que tenga, debido a que los datos pueden ser analizados en tiempo real los gráficos también lo harán solo fijando en que lapsos de tiempo el cliente requiera que se actualicen. Todas estas visualizaciones pueden ser agrupadas para formar múltiples dashboard y agruparlos por cada departamento que las PYMEs tengan, esta es otra propuesta de valor que generamos con este proyecto.

Capítulo 5 - Conclusiones y líneas de trabajo futuro

Conclusiones

Con la realización de este proyecto se pudo confirmar que el uso de software Open Source no limita la calidad de los resultados que se pueden obtener con un software licenciado similares características.

Este proyecto se demuestra que los datos de cualquier tipo de PYMEs, se pueden tratar y gestionar con la plataforma, tan solo se necesita la ayuda de un especialista de la empresa para la compresión del giro del negocio y con ello aplicar los filtros y patrones para obtener lo requerido por el cliente.

Al alojar la plataforma en la nube redujo costos de mantenimiento y respaldo de la información debido a las réplicas que oferta AWS, con ello se puede ofertar a las PYMEs la plataforma con mayor seguridad e integridad de la información.

La metodología CRISP-DM ofrece un modelo de fases a seguir flexible, ya que no es necesariamente secuencial y permite que exista una retroalimentación entre fases permitiendo la corrección de errores en caso de identificarlos. En este proyecto el uso de esta metodología nos permitió definir modelos de datos que pueden adaptarse a diferentes giros de negocio.

Líneas de trabajo futuro

Debido a las diversas características que proporciona el Stack ELK, se sugiere como trabajos futuros explotar dichas características para proporcionar a las PYMEs métricas adicionales, para que puedan tomar mejores decisiones o tener una mejor perspectiva de su negocio.

Aplicar técnicas de Machine Learning para identificar grupos de datos similares entre PYMEs, De esta forma lograr tener modelos de datos estándar que se ajusten a las necesidades u objetivos básicos de una PYME, permitiéndonos tener un enfoque más profundo en los datos y reglas de negocio propias de cada PYME.

Publicar APIs Rest con los datos procesados y transformados por la plataforma, con la finalidad de proporcionar a la PYME información de utilidad que pueda ser utilizada por la misma para la creación de nuevas aplicaciones que la ayuden a realizar la gestión de su negocio.

Bibliografía

Aalst, W. v. (2016). *Process Mining: Data Science in Action.*

AWS. (2019). *Amazon Web Services*. Obtenido de https://aws.amazon.com/es/elasticsearch-service/

AWS. (22 de 11 de 2019). *Qué es una base de datos relacional*. Obtenido de https://aws.amazon.com/es/relational-database/

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW.

Collier, K. (2011). *Agile Analytics: A Value-Driven Approach to Business Intelligence and Data Warehousing.*

Curto Díaz, J. (2010). *Introducción al Business Intelligence.*

Elastic. (15 de Noviembre de 2019). *Elastic*. Obtenido de Elastic: https://www.elastic.co/products/beats

Elastic. (2019). *Elastic Beats*. Obtenido de https://github.com/elastic/beats

Elastic Docs. (17 de Nov de 2019). *elastic*. Obtenido de https://www.elastic.co/guide/en/beats/filebeat/current/filebeat-overview.html

Gallardo Arancibia, J. A. (2009). Metodología para la Definición de Requisitos en Proyectos de Data Mining (ER-DM).

Gormley, C., & Tong, Z. (2015). Elasticsearch: The Definiteve Guide.

Guide, I. D. (2020). *IONOS*. Obtenido de https://www.ionos.es/digitalguide/hosting/cuestiones-tecnicas/bases-de-datos-relacionales/

IBM. (2016). *IBM Software.* Recuperado el 2017, de https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Luhn, H. P. (1958). A Business Intelligence System. 6.

Negash, S., & Gray, P. (2008). *Handbook on Decision Support Systems 2.*

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2014). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*.

Aalst, W. v. (2016). *Process Mining: Data Science in Action.*

AWS. (2019). *Amazon Web Services*. Obtenido de https://aws.amazon.com/es/elasticsearch-service/

AWS. (22 de 11 de 2019). *Qué es una base de datos relacional*. Obtenido de https://aws.amazon.com/es/relational-database/

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW.

Collier, K. (2011). *Agile Analytics: A Value-Driven Approach to Business Intelligence and Data Warehousing.*

Curto Díaz, J. (2010). *Introducción al Business Intelligence.*

Elastic. (15 de Noviembre de 2019). *Elastic*. Obtenido de Elastic: https://www.elastic.co/products/beats

Elastic. (2019). *Elastic Beats*. Obtenido de https://github.com/elastic/beats

Elastic Docs. (17 de Nov de 2019). *elastic*. Obtenido de https://www.elastic.co/guide/en/beats/filebeat/current/filebeat-overview.html

Gallardo Arancibia, J. A. (2009). Metodología para la Definición de Requisitos en Proyectos de Data Mining (ER-DM).

Gormley, C., & Tong, Z. (2015). Elasticsearch: The Definiteve Guide.

Guide, I. D. (2020). *IONOS*. Obtenido de https://www.ionos.es/digitalguide/hosting/cuestiones-tecnicas/bases-de-datos-relacionales/

Luhn, H. P. (1958). A Business Intelligence System. 6.

Negash, S., & Gray, P. (2008). *Handbook on Decision Support Systems 2.*

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2014). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*.